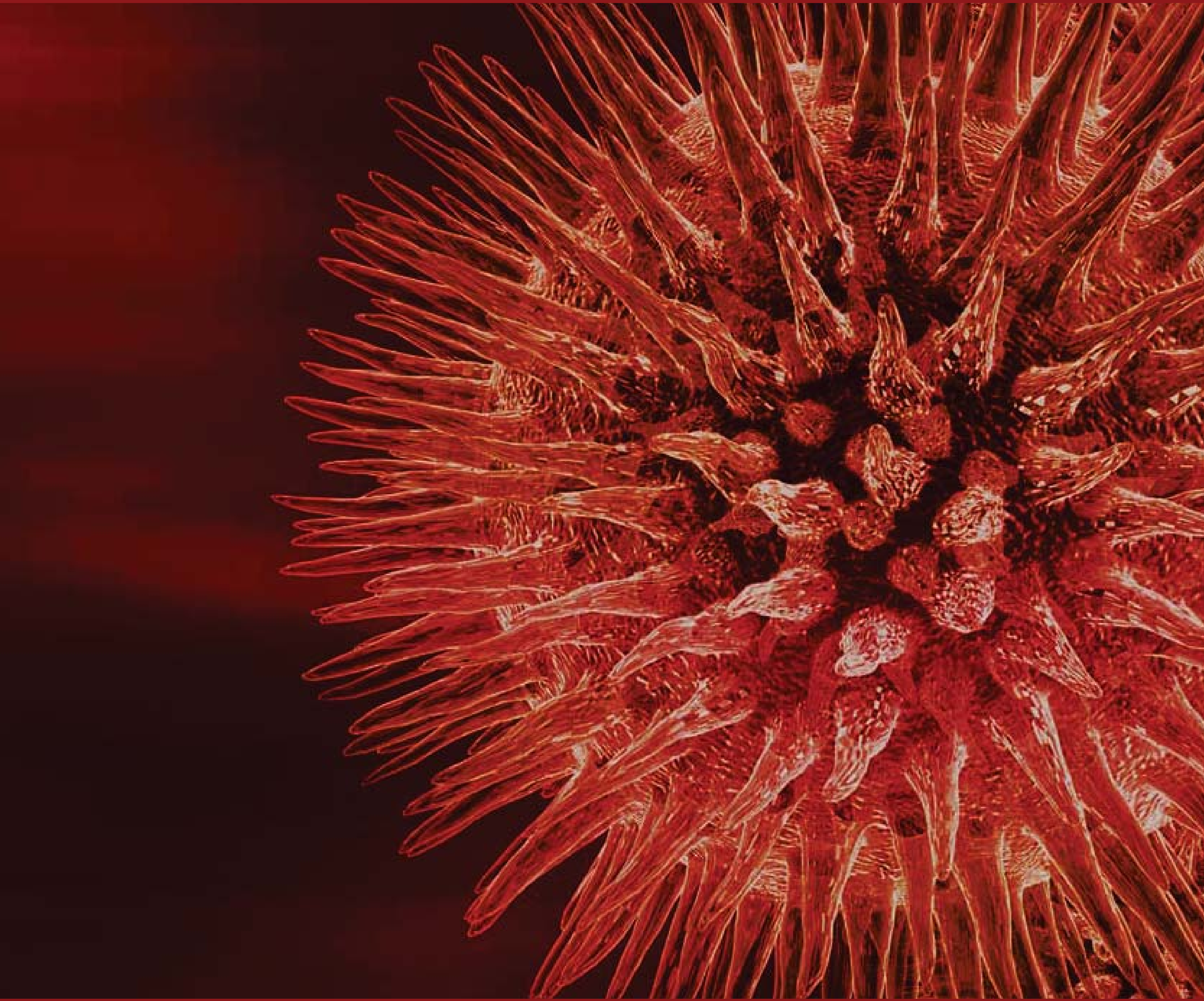
Journal of Biomedicine and Biotechnology

# Frontiers in the Convergence of Bioscience and Information Technology

Guest Editor: Daniel Howard



# Frontiers in the Convergence of Bioscience and Information Technology

Journal of Biomedicine and Biotechnology

# Frontiers in the Convergence of Bioscience and Information Technology

Guest Editor: Daniel Howard

## Editor-in-Chief

## Advisory Editors

## Associate Editors

# Contents

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 728908, 2 pages
doi:10.1155/2008/728908

*Editorial*

# Frontiers in the Convergence of Bioscience and Information Technology

**Daniel Howard**

*QinetiQ Limited, Malvern Technology Centre, St Andrew's Road, Malvern, Worcestershire WR14 3PS, UK*

Correspondence should be addressed to Daniel Howard, dr.daniel.howard@gmail.com

Received 15 June 2008; Accepted 15 June 2008

This special issue places its emphasis on the crossroads between computational modeling and the biosciences. It is ambitious but fruitful interdisciplinary research, and this special issue aims to illustrate different presentations of it. It offers those advancing knowledge in some particular area the view to related research activity and opportunities.

Many of the articles exemplify how mathematics and computer simulation inform experimentation to obtain new knowledge about nature. Some articles represent discoveries and others review, introduce, test, or trial algorithms or tools that are potentially helpful to arrive at discoveries. A few articles, however, are reversed in that they demonstrate how bioinspired algorithms or biological devices can solve hard computational challenges. Consider that manipulating chemicals may compute an answer much faster than by means of the standard silicon-based computer and within this special issue there is an article by K. Li et al. describing the principles by which such a DNA computer can compromise an encryption algorithm that is central to present day secure communications.

One might classify the articles in this special issue into three groupings: (1) those concerned with biological problems from molecular cell biology to systems biology, (2) those which apply classification algorithms in cancer diagnosis and methods of computer vision to anatomy, and finally (3) those which offer new knowledge or possibilities in bioengineering and biomedicine.

The first grouping covers Systems Biology, an exciting interdisciplinary field that marries experiments with computer simulations. Synthetic and systems biotechnology, for example, is a technology at these frontiers that aims to sequester the services of micro-organisms for the benefit of mankind. Its ability to produce foods in a vat that would otherwise take up valuable land resources with conventional agriculture will offer flexibility in food production. In their article, M.-J. Han et al. reveal how knowledge and understanding of cell physiology in the presence of oleic acid are obtained for *E. coli*. Interesting research about signaling networks in retina, as induced by light exposure in mice, is presented by J. Krishnan et al. showing marked alterations in gene expression upon light exposure. Certain transcription factors are discovered to be important for the responses to light-induced retinal loss, revealing that many of the apoptosis-related genes are up- or downregulated in this process. In their article, R. Moreno-Sanchez et al. offer a very useful review of metabolic control analysis (MCA), a tool that represents a type of engineering control theory for cell biology and when applicable MCA can help to grapple with an understanding of the complex control of the metabolic pathways.

Also in the first grouping, three articles by S. Huang and his colleagues develop the emergent and interdisciplinary field of infectomics, which is the study of infectomes encoded by the genomes of microbes and their hosts. Infectomics has potential to advance the rational strategies that will prevent and treat infectious diseases as these could require a full appreciation of the infectomes that contribute to microbial infections. There is a need to figure out how to dissect the dynamic duality relationship between symbiosis and pathogenesis in microbial infections, and advocates of this new field oppose what they see as the misguided, though current and popular, reductionist and Manichean views of the microbe-human host relationship. Another article on the topic of infectious diseases by K.-Y. Hwa et al. investigates the emerging and life-threatening infectious disease known as SARS, where there is a compelling need for the development of effective therapeutics. Their article presents interesting and potentially important findings, that molecular mimicry occurs between SARS-CoV and host proteins. They investigate how to predict those peptides that are worthy of exploration for their biological activity. The article by L. Hamel et al. provides novel insight into the field of phylogenetics with the idea of the spectra of a tree, and reviewers identified its potential to compare with phylogenetic trees across different genes and for detecting lateral gene transfer.

The first grouping of articles also covers specialist numerical algorithms for gene sequencing, genomics, and proteomics. With genomic sequencing, an error in contig assembly causes serious error proliferation, and ConPath (P.-G. Kim et al.) is a tool that can address this problem. It constructs scaffolds, ordering and orienting separate sequence contigs by exploiting the mate-pair information between contig pairs. C.-K. Chan et al. develop the growing self-organized map (GSOM) for binning (the clustering of these unassembled DNA sequences). They report improvements over the binning that combines oligonucleotide frequency and self-organizing maps (SOMs), and identify suitable training features and quantitative measures for assessing results in this area. In their article, K. Han et al. develop an algorithm to search for the highly connected subgraphs in protein interaction networks because this might be helpful to predict protein function. The article by Mi-Young Kim expounds a new approach to the very important problem of text mining of biomolecular text, and is specifically concerned with detecting gene interactions. There follow two articles in biological chemistry with computational chemistry. C.-J. Kuo et al. solve the crystal structure of H. pylori undecaprenyl pyrophosphate synthase and perform virtual screening of inhibitors from a chemical library of thousands of compounds. M. Muddassar et al. explore receptor-guided 3D-QSAR to design IGF-1R inhibitors by pursuing careful statistical analysis to interpret the models that are obtained with the help of contour maps. The article by J.-C. Lue and W.-C. Fang proposes a compact integrated microsystem solution for robust, real-time, and onsite genetic analysis. It uses a preceding VLSI differential logarithm microchip that is designed to compute the logarithm of the normalized input fluorescence signals and a succeeding VLSI artificial neural network (ANN) processor chip to analyze the processed signals from the differential logarithm stage. It is submitted that the version of ANN chosen is particularly adept at recognizing the low-fluorescence patterns.

The second grouping includes cancer diagnosis, biomedical imaging and also computational anatomy (the study of anatomical variability in health and disease via deformable templates as inspired by D'Arcy Wentworth Thompson). These fields sit in the frontiers of biosciences, image analysis, mathematics, and numerical methods. In this grouping, a number of articles implement (D. Howard et al.), evaluate (N. A. Lee et al.), or introduce (J. Kolibal et al.) techniques of image analysis to cluster, segment, or analyze images, or extract information or enrich biomedical images. For example, J. Woo et al. discuss multimodal data integration for computer-aided ablation of atrial fibrillation. Y. Park et al. analyze the interpoint dissimilarity comparisons in the hippocampus shape space to distinguish between hippocampi of subjects with three conditions (clinically depressed, high risk, and control subject), and discover the high-risk population closer in shape space to the control population than to the clinically depressed population. Additionally, they find that the left hippocampi carry more information than do the right. In their article, N. A. Lee et al. examine performance in the segmentation of high-resolution MRI subvolumes containing hippocampus, prefrontal cortex, and occipital lobe, as acquired on different scanners. They offer evidence that the alternating kernel mixture algorithm outperforms alternatives on the ten datasets considered. The availability of powerful imaging and other sensors, the availability of information technology, and the nature of modern threats point to population biometrics as a topic of enormous current and future importance. Implementing biometrics well is hard and is not yet properly understood. The methodological paper by Y. N. Shin et al. proposes a formal performance evaluation model for a biometric recognition system. They also implement face recognition systems based on the proposed model. The model seems to be useful in terms of database availability, compliance with standards, and evaluation costs. The proposed formalism may gain traction for other biometrics, and it has the potential to inform strategies for population-based biomedical imaging performance evaluation and standardization. The last article of this second grouping by J. Wichard et al. evaluates how bioinspired and computational intelligence algorithms compare in providing reliable early cancer diagnosis.

The third grouping comprises articles which cover advances in biomechanics, biophysics, and biomedicine. It includes a detailed biomechanics study by J. S. Merritt et al. on the equine distal forelimb, which is a common location of injuries related to mechanical overload. These authors combine analysis with experiment in a fairly thorough but concise way to calculate the forces in the major tendons and joint reaction from kinematic and kinetic data of walking and trotting horses. Their findings point to the importance of muscle tendon wrapping when evaluating joint loading in the distal forelimb. The article by C. Handapangoda and M. Premaratne describes a novel numerical technique for modeling optical pulse propagation in inhomogeneous scattering and absorption cross-sections through weakly scattering biological tissues. The design of implantable electronic devices to interact with the nervous system is an active field, the development of an efficient system for long-term stimulation of the optic nerve is rather timely (e.g., it is required to evaluate the long-term safety of retinal implants), and the methodological report by J. A. Zhou et al. describes the design of a suprachoroidal electrical retinal stimulator for long-term application. Finally, two articles in the third category advance novel robots (H. Sawada et al.) and haptic solutions (K.-U. Kyung et al.) that interact with the human senses and which might prove helpful to the sensorially impaired.

*Daniel Howard*

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 735101, 12 pages
doi:10.1155/2008/735101

*Research Article*

# Proteome-Level Responses of *Escherichia coli* to Long-Chain Fatty Acids and Use of Fatty Acid Inducible Promoter in Protein Production

**Mee-Jung Han,[1] Jeong Wook Lee,[1] Sang Yup Lee,[1,2] and Jong Shin Yoo[3]**

[1] *Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical & Biomolecular Engineering (BK21 Program), BioProcess Engineering Research Center, Center for Systems and Synthetic Biotechnology, and Institute for the BioCentury, Korea Advanced Institute of Science and Technology (KAIST), 335 Gwahangno, Yuseong-gu, Daejeon 305-701, South Korea*

[2] *Department of Bio and Brain Engineering and Bioinformatics Research Center, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea*

[3] *Korea Basic Science Institute, 52 Yeoeun-dong, Yuseong-gu, Daejeon 305-333, South Korea*

Correspondence should be addressed to Sang Yup Lee, leesy@kaist.ac.kr

Received 31 August 2007; Accepted 18 November 2007

Recommended by Daniel Howard

In *Escherichia coli*, a long-chain acyl-CoA is a regulatory signal that modulates gene expression through its binding to a transcription factor FadR. In this study, comparative proteomic analysis of *E. coli* in the presence of glucose and oleic acid was performed to understand cell physiology in response to oleic acid. Among total of 52 proteins showing altered expression levels with oleic acid presence, 9 proteins including AldA, Cdd, FadA, FadB, FadL, MalE, RbsB, Udp, and YccU were newly synthesized. Among the genes that were induced by oleic acid, the promoter of the *aldA* gene was used for the production of a green fluorescent protein (GFP). Analysis of fluorescence intensities and confocal microscopic images revealed that soluble GFP was highly expressed under the control of the *aldA* promoter. These results suggest that proteomics is playing an important role not only in biological research but also in various biotechnological applications.

## 1. INTRODUCTION

Exogenous fatty acids and their derivatives influence a wide variety of cellular processes including fatty acids and phospholipids synthesis, organelle inheritance, vesicle fusion, protein export and modification, enzyme activation or deactivation, cell signaling, membrane permeability, bacterial pathogenesis, and transcriptional control [1, 2]. The process governing the transport of fatty acids from environmental conditions across the membrane is distinct from the transport of hydrophilic substrates such as sugars and amino acids. In a number of cell types, the process of fatty acid transport is inducible and commensurate with the expression of specific sets of proteins [2]. In wild-type *Escherichia coli*, growth on fatty acids requires specific transport system (FadL), acyl-CoA dehydrogenase (FadD), enzymes of the $\beta$-oxidation cycle (FadA, FadB, FadE, FadF, FadG, and FadH), and glyoxylate shunt (AceA, AceB, and AceK), and these genes are negatively regulated by a transcriptional factor, FadR. Supply of long-chain fatty acids that contain 12 or more carbons results in the derepression of the genes negatively controlled by FadR but leads to the decreased expression of the genes (e.g., *fabA* and *fabB*) activated by FadR, indicating that long-chain acyl-CoA esters are the effector molecules that regulate fatty acids metabolism and thereby mediate inductions [3]. Therefore, *E. coli* cells can grow on minimal medium containing long-chain fatty acids but it cannot grow on short- and medium-chain fatty acids due to no induction of the enzymes associated with fatty acids metabolism. So far, genes involved in fatty acids metabolism (i.e., *fad* regulon) of *E. coli* have been reported at the transcriptional level by biochemical and genetic analyses [3–6]. Therefore, in this study, we looked at the effects of long-chain fatty acids at the translational level of *E. coli*.

Proteomics has changed the way to study cellular physiology. Previously, one or more proteins were chosen as models for understanding local physiological phenomena. Nowadays, proteomic studies allow researchers to identify large members of stimulons, a set of proteins whose amount or synthesis rate changes in response to a certain stimulus, and to obtain information that indicates which specific proteins should be studied further. Comparative proteome profiling under various environmental conditions also reveal new regulatory circuits and the relative abundances of protein sets at the system-wide level. Such analyses of every protein induced or repressed by the stimulus may provide the necessary information to understand a response in the cell. Furthermore, proteome profiles can prove invaluable when used in conjunction with various molecular biological tools including recombinant DNA technology [7]. For example, conditional promoters activated by the specific stimulus, such as stationary phase, pH, temperature, and nutrient limitation have been used for efficient production of heterologous proteins in bacteria [8].

In this study, proteomic studies that compared global translational differences between *E. coli* W3110 cells in the presence of glucose and oleic acid ($C_{18}$) were conducted. The present study has three goals: (i) to identify the stimulon of the oleic acid; (ii) to select target proteins from the stimulon to utilize them as the oleic acid-inducible promoter; and (iii) further to apply it for the production of recombinant proteins or other biotechnological systems.

## 2. MATERIALS AND METHODS

### 2.1. Bacterial strains and plasmids

The bacterial strains and plasmids used in this study are shown in Table 1. *E. coli* XL1-Blue was used as a strain forcloning and maintenance of plasmids. *E. coli* W3110 was used as a host strain for proteomic studies and the production of a recombinant protein. PCR primers used in this study are listed in Table 2. Primers for the amplification of the promoter regions of *aldA* and *udp* genes were designed based on the genome sequence of *E. coli* K-12 W3110 (AC_000091). The promoter region of *aldA* gene was amplified by PCR using primers 1 and 2, and was cloned into the *Eco*RV and *Eco*RI sites of pTac99A to make pAD99A (Table 1). In fact, pTac99A is a derivative of pTrc99A (Pharmacia Biotech., Uppsala, Sweden), which was constructed by replacing the *trc* promoter of pTrc99A with the *tac* promoter from pKK223-3 (Pharmacia Biotech) digested by *Pvu*II and *Eco*RI [9]. Also, the promoter region of *udp* gene was amplified by PCR using primers 3 and 4, and was cloned into the *Eco*RV and *Eco*RI sites of the high-copy-number plasmid pTac99A to make pUP99A (Table 1). Both promoters were constructed with the ribosome binding sites consisting of the AGGA sequence having an optimal distance length of 8 bases from a start codon [10].

PCR was performed in the PCR Thermal Cycler MP (Takara Shuzo Co., LTD., Shiga, Japan) using the Expand High Fidelity PCR System (Roche Molecular Biochemicals, Mannheim, Germany). DNA sequencing was carried out using the Bigdye terminator cycle sequencing kit (Perkin-Elmer Co., Boston, Mass, USA), Taq polymerase and the ABI Prism 377 DNA sequencer (Perkin-Elmer Co., Mass, USA). All DNA manipulations were carried out according to standard procedures [11].

### 2.2. Cell growth conditions and analytical procedure

Cells were cultivated at 37°C and 250 rpm in 100 mL of Luria-Bertani (LB) medium (10 g/L of tryptone, 5 g/L of yeast extract, and 5 g/L of NaCl), or R/2 medium plus 10 g/L glucose or 5 g/L oleic acid (Daejung Chemicals & Metals Co., Gyeonggi-do, Korea) as a carbon source. The R/2 medium (pH 6.8) contains per liter: 2 g of $(NH_4)_2HPO_4$, 6.75 g of $KH_2PO_4$, 0.85 g of citric acid, 0.7 g of $MgSO_4 \cdot 7H_2O$, and 5 mL of a trace metal solution. The trace metal solution contains per liter of 5 M HCl : 10 g of $FeSO_4 \cdot 7H_2O$, 2.25 g of $ZnSO_4 \cdot 7H_2O$, 1 g of $CuSO_4 \cdot 5H_2O$, 0.5 g of $MnSO_4 \cdot 5H_2O$, 0.23 g of $Na_2B_4O_7 \cdot 10H_2O$, 2 g of $CaCl_2 \cdot 2H_2O$, and 0.1 g of $(NH_4)_6MO_7O_{24}$. For the cultivation of recombinant *E. coli* strains, ampicillin (Ap, 50 $\mu$g/mL) was added. Cell growth was monitored by measuring the absorbance at 600 nm ($OD_{600}$; DU Series 600 Spectrophotometer, Beckman, Fullerton, Calif, USA). At an $OD_{600}$ of 0.7 or 1.2, isopropyl-$\beta$-D-thiogalactopyranoside (IPTG, Sigma Chemical Co., St. Louis, Mo, USA) was added at a final concentration of 1 mM. For induction by oleic acid, the defined medium supplemented with 10 g/L glucose was changed into the medium plus 5 g/L oleic acid after cells were collected by centrifugation at the same $OD_{600}$ of 0.7 or 1.2. Then, cells were further cultivated for 5, 10, and 20 hours, and harvested by centrifugation at 3,500 × g for 5 minutes at 4°C. Protein samples were analyzed by electrophoresis on 12% (w/v) sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) as described by Laemmli [12]. The gels were stained with Coomassie brilliant blue R250 (Bio-Rad, Hercules, Calif, USA), and the protein bands were quantified by a GS-710 Calibrated Imaging Densitometer (Bio-Rad).

### 2.3. Two-dimensional gel electrophoresis (2DE)

Proteome analysis was performed by 2DE using the IPGphor IEF system (GE Healthcare, Chalfont St. Giles, UK) and Protean II xi Cell (Bio-Rad) as described previously [13]. In brief, *E. coli* W3110 cells grown in the presence of glucose and oleic acid were harvested at the exponential and stationary phases, respectively, by centrifugation for 5 minutes at 3,500 × g and 4°C, and washed four times with low-salt washing buffer. The pellet was then resuspended in 600 $\mu$L of a buffer containing 10 mM Tris-HCl (pH 8.0), 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.1% (w/v) SDS, and 1% (v/v) cocktail protease inhibitor (Complete Mini EDTA-free; Roche Diagnostics GmbH, Germany). One $\mu$L of this sample was mixed with 60 $\mu$L of a solution consisting of 8 M urea, 4% (w/v) CHAPS, 40 mM Tris, 65 mM DTT, and a trace of bromophenol blue. Proteins (200 $\mu$g) quantified by Bradford assay [14] were resuspended in 350 $\mu$L of IEF denaturation buffer composed of 8 M urea, 2% (w/v) CHAPS, 20 mM DTT, and 0.8% (v/v) IPG buffer (pH 3–10 NL;

Table 1: Bacterial strains and plasmids used in this study.

| Strain or plasmid | Relevant characteristics | Reference or source |
|---|---|---|
| *E. coli strains* | | |
| XL1-Blue | *recA1, endA1, gyrA96, thi, hsdR17, suppE44, relA1, l*$^-$*, lac*$^-$*, F'[proAB lacI*$^q$ *lacZ* Δ*M15, Tn10 (tet)*$^r$*]* | Stratagene[a] |
| W3110 | *F*$^-$ *mcrA mcrB IN(rrnD*$^-$*rrnE)1λ*$^-$ | KCTC[b] |
| *Plasmids* | | |
| pTac99A | pTrc99A derivative; *tac* promoter, cloning vehicle; Ap$^r$ | Park and Lee [9] |
| pAD99A | pTac99A derivative; aldehyde dehydrogenase (*aldA*) promoter; Ap$^r$ | This study |
| pUP99A | pTac99A derivative; uridine phosphorylase (*udp*) promoter; Ap$^r$ | This study |
| pGFPuv | Ap$^r$, *lac* promoter, *gfp* | Clontech[c] |
| pAD99GFP | pAD99A derivative; *gfp* | This study |
| pTac99GFP | pTac99A derivative; *gfp* | This study |
| pUP99GFP | pUP99A derivative; *gfp* | This study |

[a] Stratagene Cloning System (La Jolla, Calif, USA).
[b] Korean Collection for Type Cultures, (Daejeon, Korea).
[c] BD Biosciences Clontech (Palo Alto, Calif, USA).

Table 2: List of primers used in PCR experiments.

| Primer | Primer sequence[a] | Gene to be amplified | Template |
|---|---|---|---|
| Primer 1 | aaaaccgtt**gatatc**tttgcaaacgggcatgactcctgactttt | *aldA* promoter | *E. coli* W3110 chromosome |
| Primer 2 | aaaaccgtt**gaattc**ctcctgtgatttatatgtttgttttc | | |
| Primer 3 | aaaaccgtt**gatatc**tgcagaatgaagggtgatttatgtgatttg | *udp* promoter | *E. coli* W3110 chromosome |
| Primer 4 | aaaaccgtt**gaattc**ctcctctgtgaatcggtttagtcaga | | |
| Primer 5 | g**gaattc**atgagtaaaggagaagaacttt | GFP | pGFPuv |
| Primer 6 | ccc**aagctt**ttatttgatgagctcatcc | | |

[a] Restriction enzyme sites are shown in bold.

GE Healthcare). The samples were carefully loaded on the IPG strips (18 cm, pH 3–10 NL; GE Healthcare). The loaded IPG strips were rehydrated for 12 hours and focused at 20°C for 15 minutes at 250 V, followed by 8,000 V until a total of 60,000 V·h was reached. The strips were equilibrated in two equilibration buffers as described previously [15] and then placed on 12% (w/v) SDS-PAGE gels prepared by the standard protocol [12]. Protein spots were visualized using a silver staining kit (GE Healthcare), and the stained gels were scanned by a GS-710 Calibrated Imaging Densitometer (Bio-Rad). ImageMaster 2D Platinum Software (version 5.0; GE Healthcare) was used to identify spots, to match gels, and to quantify spot densities on a volume basis (i.e., integration of spot optical intensity over the spot area).

### 2.4. Fractionation of outer membrane proteins

Culture broth (3 mL) was centrifuged at 3,500 × g for 5 minutes at 4°C, and the pellet was washed with 1 mL of 10 mM $Na_2HPO_4$ buffer (pH 7.2), followed by centrifugation at 3,500 × g for 5 minutes at 4°C. The cell pellet was resuspended in 0.5 mL of 10 mM $Na_2HPO_4$ buffer (pH 7.2). Crude extracts of *E. coli* cells were prepared by five cycles of sonication (each for 15 seconds at 20% of maximum output; High-intensity ultrasonic liquid processors; Sonics & Material Inc., Newtown, Conn, USA). Partially disrupted cells were first removed by centrifugation of sonicated samples at 12,000 × g for 2 minutes at room temperature. Membrane proteins and lipid layers were isolated by centrifugation at 12,000 × g for 30 minutes at 4°C, followed by resuspension in 0.5 mL of 0.5% (w/v) sarcosyl in 10 mM $Na_2HPO_4$ buffer (pH 7.2). After incubation at 37°C for 30 minutes, the insoluble pellet containing membrane proteins was obtained by centrifugation at 12,000 × g for 30 minutes at 4°C. Membrane proteins were obtained by washing the insoluble pellet with 10 mM $Na_2HPO_4$ buffer (pH 7.2), followed by resuspending in 50 $\mu$L of Tris-EDTA buffer (pH 8.0).

### 2.5. Protein identification by LC-MS/MS analysis

Samples for the MS/MS analysis were prepared as described previously [16]. Briefly, protein spots were excised and destained by incubating in 30 mM potassium ferricyanide and 65 mM sodium thiosulfate for 10 minutes. Gel pieces were washed in Milli-Q water until they became colorless and transparent, and then vacuum-dried. These pieces were proteolysed with 0.02 $\mu$g/$\mu$L of modified trypsin (Promega, Madison, Wis, USA) in 40 mM ammonium bicarbonate for overnight at 37°C. Tryptic peptides (10 $\mu$L aliquots) were analyzed by a nano-LC/MS system consisting of an Ultimate HPLC system (LC Packings, Amsterdam, Netherlands) and a quadrupole-time-of-flight (Q-TOF) MS (Micromass, Manchester, UK) equipped with a nano-ESI source as described previously [15]. The MASCOT search server (version 1.8; http://www.matrixscience.com) was used for the identification of protein spots by querying sequence of the tryptic peptide fragments. Reference databases used for the identification of target proteins were UniProt Knowledgebase (Swiss-Prot and TrEMBL; http://kr.expasy.org) and NCBI (http://www.ncbi.nlm.nih.gov).

### 2.6. Fluorescence microscopy and intensity of GFP

For fluorescence imaging, cells were harvested by centrifugation for 5 minutes at 3,500 × g and 4°C, washed with and resuspended in phosphate-buffered saline (PBS) solution. The samples were mounted on microscopic slide glasses and examined by confocal microscopy (Carl Zeiss, Jena, Germany). Photographs were taken with a Carl Zeiss LSM 410 instrument. Samples were excited by a 364-nm argon laser, and images were filtered by a longpass 505-nm filter. Three-dimensional images were constructed from 5–10 serial images (each 1-Am thick) made by automatic optical sectioning. Fluorescence intensities were measured at 395 nm (excitation) and 509 nm (emission) using the SpectraMax M2 multi-detection system (Molecular Devices, Sunnyvale, Calif, USA), and a 96-well black and clear flat-bottom plate (Coastar, Los Angeles, Calif, USA).

## 3. RESULTS AND DISCUSSION

### 3.1. Proteome analysis

To understand physiological changes triggered by the long-chain fatty acid, we analyzed the proteome profiles of *E. coli* K-12 W3110 grown in the presence of glucose and oleic acid, respectively. The final concentration of cells cultured in oleic acid as a carbon source was 4-fold higher than that of cells grown in glucose, although the former took a longer lag-period for induction of the *fad* regulon (see Figure 1). Samples of proteome were taken at the exponential and stationary phases in two different media (see Figure 1): when the $OD_{600}$ of *E. coli* reached 0.57 and 1.25 in the presence of glucose, named G1 and G2, respectively; and when the $OD_{600}$ reached 0.56 and 5 in the presence of oleic acid, named O1 and O2, respectively. The proteome profiles of the four samples, G1, G2, O1, and O2, were analyzed by 2D PAGE using a strip of 3–10 p*I* range and 12% polyacrylamide gel for subsequent comparisons (see Figure 2). The overall profiles of whole cellular proteins were reproducible. From over 2,000 spots on each 2D gel shown in Figure 2, we identified 92 proteins by comparing with our in-house *E. coli* proteome database or by conducting LC-MS/MS analysis. Functions and fold changes of individual proteins are shown in Table 3.

The outer membrane proteins were enriched by fractionation, and separated on 12% SDS-PAGE (see Figure 3). Membrane proteins are typically difficult to be resolved in the IEF denaturation buffer used commonly for 2D gels because of their hydrophobic property. The highly abundant porin, OmpF whose expression level was regulated by osmolarity [17], was observed at the exponential phase of *E. coli* in the presence of glucose. As expected, the long-chain fatty acid transporter protein, FadL, was newly synthesized in the presence of oleic acid. This result proves that *E. coli* requires the specific transport system (*fadL*) on the growth of fatty acids.

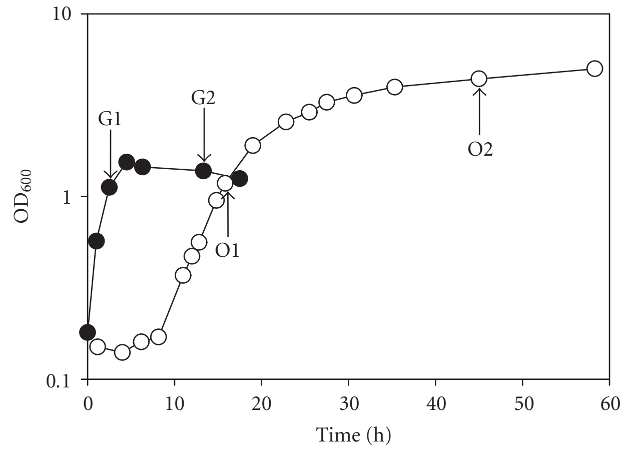


Figure 1: Time profiles of the concentrations of *E. coli* cells. The cell densities ($OD_{600}$) of *E. coli* W3110 in the presence of glucose (•) or oleic acid (∘) are shown. Gl, G2, O1, and O2 are the sampling points for proteome analyses.

### 3.2. Identification of the proteins stimulated by oleic acid

To examine the influence of oleic acid on the proteome profile variation, we compared the proteomes obtained from the exponential phase (O1 versus G1) and the stationary phase (O2 versus G2), as shown in Table 3. At the exponential growth phase, the levels of 41 identified proteins were altered in the presence of oleic acid. Among them, 9 proteins including AldA, Cdd, FadA, FadB, FadL, MalE, RbsB, Udp, and YccU were newly synthesized in response to oleic acid, while GapA (the fragment), hypothetical protein YfdX, and two unidentified proteins were not detectable. As expected, the levels of proteins involved in fatty acid degradation (FadA and FadB), long-chain fatty acid transport system (FadL), glyoxylate shunt (AceA). and TCA cycle (Mdh, SdhA, SucC, and SucD) were significantly increased to replenish the dicarboxylic acid intermediates consumed in amino acid biosynthesis. Particularly, isocitrate lyase (AceA) in the *aceBAK* operon was significantly synthesized by more than five folds in the presence of oleic acid, making it the most abundant protein. Concurrently, there were decreased levels of proteins involved in the biosynthesis of fatty acids (FabD and FabE) and amino acids (AroG, LeuC, and SerC). These results showed that the variation patterns of most proteins identified as a *fad* regulon were in agreement with their corresponding transcriptional levels previously reported [2–6].

Furthermore, the growth of *E. coli* on oleic acid involves a significant contribution of the pyrimidine salvage pathway (Cdd and Udp) and specific binding-protein-dependent transport system (MalE and RbsB) because the levels of these proteins highly increased by oleic acid. The salvage pathway of *E. coli* functions to reutilize free bases and nucleosides

Table 3: Proteins identified from 2DE.

| Spot no. | Protein name | Method for identity | Accession no. | $p^I/Mw^a$ (kDa) | Protein description | Fold change[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | O1/G1 | O2/G2 |
| 1 | AcnB | Gel match | P36683 | 5.24/75.9 | Aconitate hydratase 2 | Δ | Δ |
| 2 | AsnS | Gel match | P17242 | 5.64/92.8 | Asparaginyl-tRNA synthetase | — | — |
| 3 | AceF | Gel match | P06959 | 5.01/77.5 | Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex | ∇ | — |
| 4 | DnaK | Gel match | P04475 | 4.81/69.6 | Chaperone protein DnaK | ∇ | ∇ |
| 5 | PtsI | Gel match | P08839 | 4.78/59.8 | Phosphoenolpyruvate-protein phosphotransferase | ∇ | 3×∇ |
| 6 | AldA | MS/MS | P25553 | 5.07/52.2 | Aldehyde dehydrogenase A | Appeared | Appeared |
| 7 | MopA (GroEL) | Gel match | P06139 | 4.85/56.7 | 60 kDa chaperonin (GroEL protein) | ∇ | ∇ |
| 8 | Tig | Gel match | P22257 | 4.83/51.0 | Trigger factor (TF) | — | — |
| 9 | HtpG | Gel match | P10413 | 5.06/65.6 | Chaperone protein HtpG (Heat shock protein HtpG) | — | — |
| 10 | AtpD | Gel match | P00824 | 4.90/47.7 | ATP synthase beta chain | — | — |
| 11 | Icd (IcdA) | Gel match | P08200 | 5.02/46.0 | Isocitrate dehydrogenase | — | Δ |
| 12 | AceA1 | MS/MS | P05313 | 5.19/44.1<br>5.12/44.1 | Isocitrate lyase | 5×Δ | 7×Δ |
| | AceA2 | MS/MS | | 5.01/33.8 | Isocitrate lyase fragment | 6×Δ | 7×Δ |
| 13 | GlnA | Gel match | P06711 | 5.25/53.8 | Glutamine synthetase | — | — |
| 14 | IlvC | Gel match | P05793 | 5.26/52.0 | Ketol-acid reductoisomerase | — | — |
| 15 | GlpK | Gel match | P08859 | 5.30/50.6 | Glycerol kinase (Glycerokinase) | — | — |
| 16 | Eno | Gel match | P08324 | 5.34/46.5<br>5.29/46.2 | Enolase (2-phosphoglycerate dehydratase) | — | ∇ |
| 17 | TufA (EF-Tu) | Gel match | P02990 | 5.32/44.6 | Elongation factor Tu (EF-Tu) | — | — |
| 18 | FabD (TfpA) | Gel match | P25715 | 5.37/44.8 | Malonyl CoA-acyl carrier protein transacylase | ∇ | — |
| 19 | LeuC | Gel match | P30127 | 5.42/44.4<br>5.95/51.6 | 3-isopropylmalate dehydratase large subunit | ∇ | — |
| 20 | FadB | MS/MS | P21177 | 5.84/79.5 | Fatty oxidation complex alpha subunit | Appeared | Appeared |
| 21 | SdhA | Gel match | P10444 | 5.74/63.7 | Succinate dehydrogenase flavoprotein subunit | — | — |
| 22 | OppA | Gel match | P23843 | 5.93/56.1 | Periplasmic oligopeptide-binding protein | ∇ | — |
| 23 | TrpD | Gel match | P00904 | 6.08/55.9 | Anthranilate synthase component II; Anthranilate | — | — |
| 24 | GuaB (GuaR) | Gel match | P06981 | 6.01/55.0 | Inosine-5′-monophosphate dehydrogenase | ∇ | ∇ |
| 25 | AtpA | Gel match | P00822 | 5.84/53.1 | ATP synthase alpha chain | — | Δ |
| 26 | DppA | Gel match | P23847 | 5.69/52.1 | Periplasmic dipeptide transport protein | — | — |
| 27 | GlyA | Gel match | P00477 | 6.04/45.9<br>5.94/46.1 | Serine hydroxymethyltransferase (Serine methylase) | — | — |
| 28 | CarA (PyrA) | Gel match | P00907 | 5.91/44.0 | Carbamoyl-phosphate synthase small chain | — | — |
| 29 | Unknown | MS/MS | — | — | — | Disappeared | — |

Table 3: Continued.

| Spot no. | Protein name | Method for identity | Accession no. | $p^I$/Mw[a] (kDa) | Protein description | Fold change[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | O1/G1 | O2/G2 |
| 30 | Unknown | MS/MS | — | — | — | Disappeared | Disappeared |
| 31 | FadA | MS/MS | P21151 | 6.31/40.9 | Fatty oxidation complex beta subunit | Appeared | Appeared |
| 32 | Fba | Gel match | P11604 | 5.55/40.6 | Fructose-bisphosphate aldolase class II | — | — |
| 33 | SerC (PdxF) | Gel match | P23721 | 5.34/40.2 | Phosphoserine aminotransferase | ∇ | ∇ |
| 34 | SucC | Gel match | P07460 | 5.30/42.3 | Succinyl-CoA synthetase beta chain | Δ | Δ |
| 35 | LivJ | MS/MS | P02917 | 5.28/41.9 | Leu/Ile/Val-binding protein | — | — |
| 36 | Pgk | Gel match | P11665 | 5.07/41.9<br>5.02/41.7 | Phosphoglycerate kinase | — | ∇ |
| 37 | MalE | MS/MS | P02928 | 5.08/41.1 | Maltose-binding periplasmic protein | Appeared | Appeared |
| 38 | LivK | Gel match | P04816 | 5.00/41.4 | Leucine-specific binding protein | — | — |
| 39 | RfaD (HtrM) | Gel match | P17963 | 4.85/36.8 | ADP-L-glycero-D-manno-heptose-6-epimerase | — | — |
| 40 | PotD | Gel match | P23861 | 4.77/35.8 | Spermidine/putrescine-binding periplasmic protein | — | — |
| 41 | TalB | Gel match | P30148 | 5.01/35.8 | Transaldolase B | — | ∇ |
| 42 | Tsf (EF-Ts) | Gel match | P02997 | 5.15/33.6 | Elongation factor Ts (EF-Ts) | — | ∇ |
| 43 | Mdh | MS/MS | P06994 | 5.55/35.5 | Malate dehydrogenase | Δ | — |
| 44 | CysK | Gel match | P11096 | 5.81/36.0 | Cysteine synthase A | — | — |
| 45 | ManX (PtsL) | Gel match | P08186 | 5.17/26.1 | PTS system, mannose-specific IIAB component | ∇ | — |
| 46 | Unknown | MS/MS | — | — | — | 2×∇ | 3×∇ |
| 47 | AroG | Gel match | P00886 | 6.12/39.4 | Phospho-2-dehydro-3-deoxyheptonate aldolase | 2×∇ | ∇ |
| 48 | Sbp | Gel match | P06997 | 6.49/49.5 | Sulfate-binding protein | 2×∇ | ∇ |
| 49 | GapA | Gel match | P06977 | 6.58/36.3 | Glyceraldehyde 3-phosphate dehydrogenase A | — | — |
| 50 | PyrB | Gel match | P00479 | 6.13/35.3 | Aspartate carbamoyltransferase catalytic chain | — | — |
| 51 | FkpA | Gel match | P45523 | 7.08/33.2 | FKBP-type peptidyl-prolyl *cis-trans* isomerase FkpA | — | — |
| 52 | SucD | MS/MS | P07459 | 6.31/29.6 | Succinyl-CoA synthetase alpha chain | Δ | Δ |
| 53 | GapA | MS/MS | P06977 | 6.58/23.0 | No. 49 fragment | Disappeared | Disappeared |
| 54 | GlnH | MS/MS | P10344 | 6.87/24.9 | Glutamine-binding periplasmic protein | — | — |
| 55 | SodA | Gel match | P00448 | 6.44/22.9 | Superoxide dismutase [Mn] (MnSOD) | — | — |
| 56 | RbsB | MS/MS | P02925 | 5.92/29.1 | D-ribose-binding periplasmic protein | Appeared | Appeared |
| 57 | Udp | Gel match | P12758 | 5.86/27.9 | Uridine phosphorylase (UDRPase) | Appeared | Appeared |
| 58 | YadK | Gel match | P37016 | 5.55/28.4 | Protein YadK | — | — |
| 59 | TpiA (Tpi) | Gel match | P04790 | 5.57/26.9 | Triosephosphate isomerase | — | — |

Table 3: Continued.

| Spot no. | Protein name | Method for identity | Accession no. | p$^I$/Mw$^a$ (kDa) | Protein description | Fold change[b] | |
|---|---|---|---|---|---|---|---|
| | | | | | | O1/G1 | O2/G2 |
| 60 | Cdd | MS/MS | P13652 | 5.08/31.5 | Cytidine deaminase | Appeared | Appeared |
| | | MS/MS | | 5.42/31.5 | | Appeared | Appeared |
| 61 | TrpA | Gel match | P00928 | 5.30/28.7 | Tryptophan synthase alpha chain | — | — |
| 62 | SspA (Ssp) | Gel match | P05838 | 5.24/26.6 | Stringent starvation protein A | — | — |
| 63 | HisJ | Gel match | P39182 | 5.05/28.6 | Histidine-binding periplasmic protein | — | — |
| 64 | FliY | Gel match | P39174 | 5.01/26.2<br>5.11/25.8 | Cystine-binding periplasmic protein | ∇ | 3×∇ |
| 65 | HdhA (HsdH) | Gel match | P25529 | 5.17/25.0 | 7-alpha-hydroxysteroid dehydrogenase | ∇ | ∇ |
| 66 | Upp (UraP) | Gel match | P25532 | 5.29/23.8 | Uracil phosphoribosyltransferase | — | — |
| 67 | GrpE | Gel match | P09372 | 4.68/25.5 | GrpE protein (HSP-70 cofactor) | — | 2×∇ |
| 68 | AccB (FabE) | Gel match | P02905 | 4.57/22.0 | Biotin carboxyl carrier protein of acetyl-CoA carboxylase | 2×∇ | 2×∇ |
| 69 | YfdX | MS/MS | P76520 | 5.38/23.0 | Protein yfdX | Disappeared | Disappeared |
| 70 | AhpC | Gel match | P26427 | 5.01/21.5 | Alkyl hydroperoxide reductase C22 protein | ∇ | ∇ |
| 71 | Crr | Gel match | P08837 | 4.57/20.0<br>4.68/18.9 | PTS system, glucose-specific IIA component | — | — |
| 72 | DksA | Gel match | P18274 | 4.90/18.7 | DnaK suppressor protein | — | ∇ |
| 73 | AroK | Gel match | P24167 | 5.30/17.9 | Shikimate kinase I | — | — |
| 74 | SodB | Gel match | P09157 | 5.53/22.1 | Superoxide dismutase [Fe] | 2×∇ | ∇ |
| 75 | PpiB | Gel match | P23869 | 5.51/17.7 | Peptidyl-prolyl cis-trans isomerase B | — | ∇ |
| 76 | RplI | Gel match | P02418 | 6.20/19.8<br>6.17/15.7 | 50S ribosomal protein L9 | ∇ | ∇ |
| 77 | YbdQ | Gel match | P39177 | 6.08/15.5 | Unknown protein from 2D-page | — | — |
| 78 | RbfA | Gel match | P09170 | 6.00/15.6 | Ribosome-binding factor A | — | — |
| 79 | RplU | Gel match | P02422 | 6.71/10.3 | 50S ribosomal protein L21 | 2×∇ | 3×∇ |
| 80 | Hns | Gel match | P08936 | 5.45/15.6 | DNA-binding protein H-NS (Histone-like protein HLP-II) | — | Δ |
| 81 | Ndk | Gel match | P24233 | 5.59/15.2 | Nucleoside diphosphate kinase (NDP kinase) | — | — |
| 82 | AtpC | Gel match | P00832 | 5.48/14.8 | ATP synthase epsilon chain | — | — |
| 83 | RpsF | Gel match | P02358 | 5.31/15.8<br>5.15/15.8<br>5.26/15.8 | 30S ribosomal protein S6 | — | — |
| 84 | Bcp | Gel match | P23480 | 5.02/15.8 | Bacterioferritin comigratory protein | ∇ | 2×∇ |
| 85 | GreA | Gel match | P21346 | 4.68/15.9 | Transcription elongation factor GreA | 2×∇ | Disappeared |
| 86 | GroES (MopB) | Gel match | P05380 | 5.15/15.6 | 10 kDa chaperonin (GroES protein) | — | — |
| 87 | YfiD | MS/MS | P33633 | 5.09/14.3 | Protein YfiD | — | — |

Table 3: Continued.

| Spot no. | Protein name | Method for identity | Accession no. | p$^I$/Mw$^a$ (kDa) | Protein description | Fold change$^b$ | |
|---|---|---|---|---|---|---|---|
| | | | | | | O1/G1 | O2/G2 |
| 88 | UspA | Gel match | P28242 | 5.14/15.1 | Universal stress protein A | — | — |
| 89 | YjgF | Gel match | P39330 | 5.29/13.0 | Protein YjgF | ∇ | ∇ |
| 90 | TrxA (TsnC) | Gel match | P00274 | 4.67/11.5 | Thioredoxin 1 | — | — |
| 91 | HdeB | Gel match | P26605 | 4.85/11.2 | Protein HdeB (10K-L protein) | — | 3×∇ |
| 92 | YccU | MS/MS | P75874 | 6.72/14.7 | Protein YccU; Predicted CoA-binding protein | Appeared | Appeared |

[a] Unit of the molecular weight (MW) is kDa.
[b] Fold change: 0 ~0.3-fold, 3×∇; 0.3 ~0.5-fold, 2×∇; 0.5 ~0.6-fold, ∇; 1.5-fold, Δ; 2-fold, 2×Δ; fold change, fold number×Δ.

produced intracellularly from nucleotide turnover [18]. Also, the pyrimidine salvage pathway has been reported to recycle the pentose moieties of exogenous nucleosides to use them as carbon and energy sources and the amino groups of cytosine compounds as a nitrogen source. The D-ribose-binding periplasmic protein, RbsB in the *rbsACBK* operon, mediates the entry of D-ribose across the cell membrane in the form of D-ribose 5-phosphate, which is an intermediate of the pentose phosphate cycle [19]. Therefore, long-chain fatty acids seem to influence the status of the pyrimidine salvage pathway and its associated transport system.

Interestingly, aldehyde dehydrogenase (AldA), which oxidize diverse aldehydes throughout the cellular metabolism, received a special attention in this study because of its possible applications in gene expression system with oleic acid as an inducer. It has been reported that the expression of *aldA* gene was induced on growth on fucose, rhamnose, arabinose, glutamate, or 2-oxoglutarate during aerobic condition, while that is repressed by glucose [20]. Our observation found in this study demonstrated that oleic acid is another inducer of the *aldA* gene. Its application as an inducible promoter is demonstrated in the next two sections.

At the stationary phase, the levels of 45 identified proteins were altered in response to oleic acid (Table 3). Sixteen proteins including AceA, AcnB, AldA, AtpA, Cdd, FadA, FadB, FadL, Hns, Icd, MalE, RbsB, SucC, SucD, Udp, and YccU were significantly increased or newly synthesized, while 29 proteins were reduced or disappeared on 2D gels. Although the variations of proteins at the stationary phase were more complex, most proteins with altered levels on 2D gels at this phase showed patterns similar to those at the exponential growth phase.

### 3.3. Construction of the expression system with oleic acid-inducible promoters

Among proteins highly inducible by oleic acid, we selected two target proteins, AldA and Udp according to the following two criteria for their utilization as promoters: (i) they are only induced in the presence of oleic acid to be strictly controlled; (ii) they are strongly and highly expressed for the enhanced bioproducts production. AldA and Udp were synthesized in response to oleic acid with relatively high abundance from the exponential to stationary phases, and were not synthesized in the presence of glucose, suggesting that the native promoters of these proteins could be used as an oleic acid-inducible promoter in *E. coli* W3110.

For the construction of the expression systems controlled by *aldA* or *udp* promoter, the promoter regions of these genes were amplified as described in Section 2. The representative schematic plasmid map under the control of *aldA* promoter is illustrated in Figure 4. This is a high-copy-number plasmid with replication origin of pBR322 (ATCC 37017).

### 3.4. Comparison between the expression efficiency of oleic acid- and IPTG-inducible promoters

Various cultivation strategies employing different host strains and expression systems have been employed for the production of recombinant proteins [10, 21]. One of the most popular approaches is the use of different promoters to regulate expression levels [10]. In *E. coli*, many inducible promoters have been developed, which can be induced by various mechanisms such as temperature upshifting, pH fluctuation, nutrient starvation, and addition of chemical inducers. Among these inducible systems, T7 or *lac*-based promoters (*tac*, *trc*, *lac*, *lac*UV5-T7 hybrid, etc.), which can be effectively induced by the addition of IPTG, are the most frequently used ones.

In order to evaluate the effectiveness of oleic acid-inducible promoters discovered in this study, we chose the IPTG-inducible *tac* promoter as a control. Green fluorescent protein (GFP) from the jellyfish *Aequorea Victoria* was employed as a model recombinant protein to examine its expression under the control of *aldA* or *udp* promoter. For the induction of GFP by oleic acid, the defined medium supplemented with glucose was transferred into oleic acid medium at the $OD_{600}$ of 0.7 or 1.2. For the control, cells harboring the plasmid containing *tac* promoter were added with 1 mM IPTG at the same values of $OD_{600}$ in the defined medium supplemented with glucose. After induction by IPTG or oleic acid, cells were further cultured, harvested at each time, and analyzed by 12% SDS-PAGE (see Figure 5). When GFP is induced at the $OD_{600}$ of 1.2 in recombinant *E. coli* W3110 harboring pTac99GFP, pAD99GFP, and pUP99GFP, its contents were approximately 27%, 42%, and 25% of the total proteins at 10 hours, and 28%, 50%, and 25% at 20 hours, respectively. The GFP content induced by *aldA* promoter was

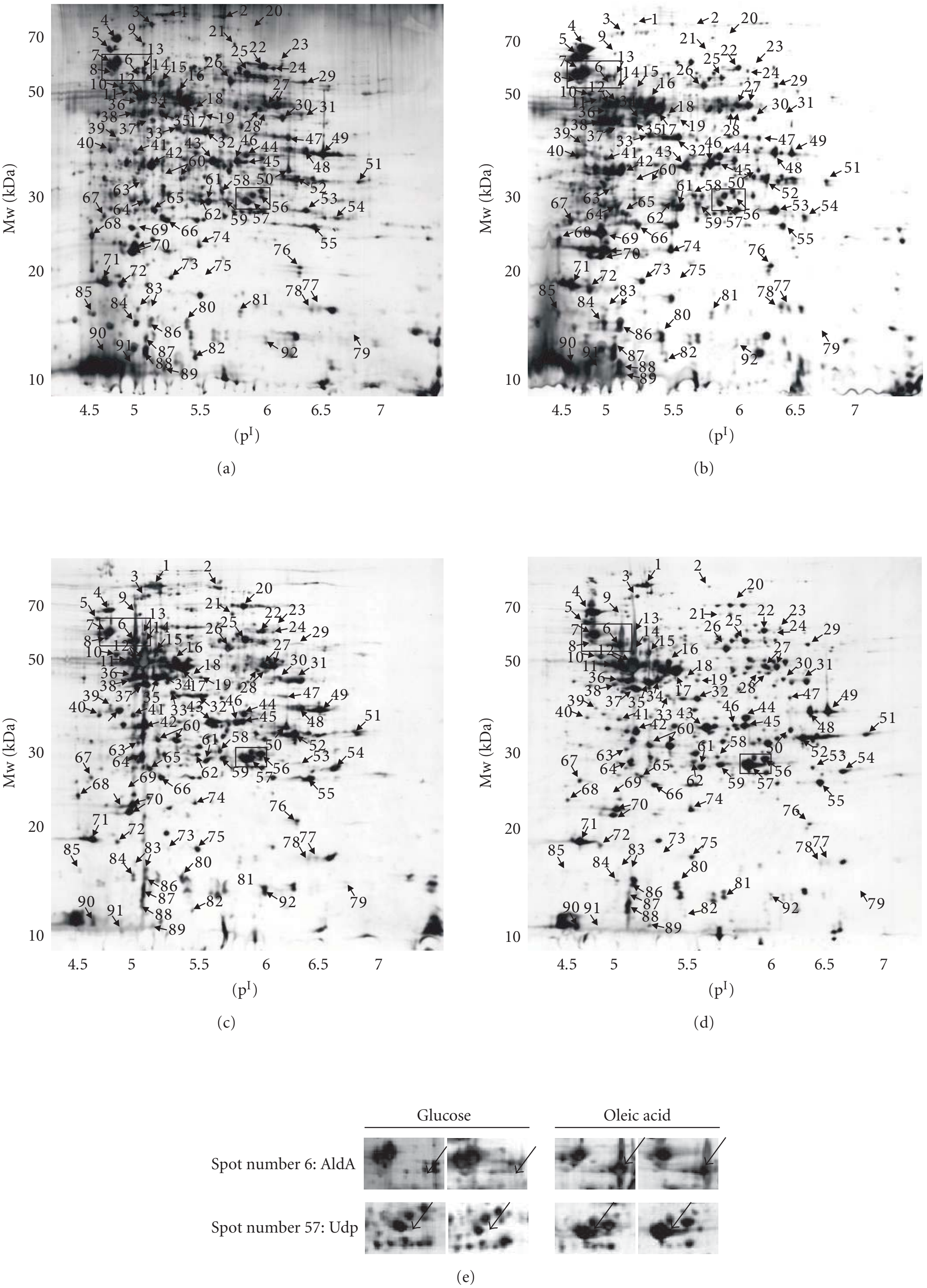


Figure 2: The 2DE maps of *E. coli* W3110 cells at the exponential (left panels; A, C) and stationary phases (right panels; B, D) in the presence of glucose (A, B) and oleic acid (C, D), respectively. Identified proteins shown in numbers are listed in Table 3. Boxes further highlight specific corresponding regions of the 2D gel images, which are compared at higher resolution in the bottom of (E). Arrow lines indicate individual spots of AldA and Udp.

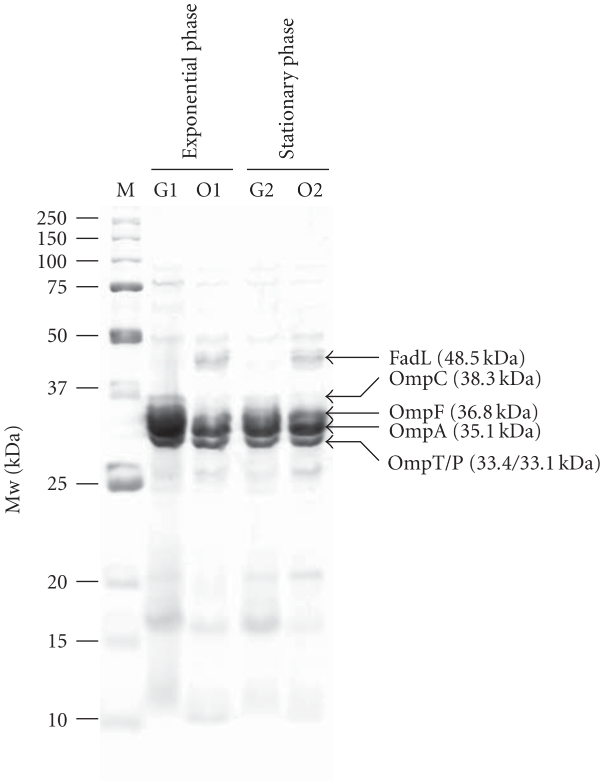


Figure 3: SDS-PAGE of the outer membrane proteins taken from samples of proteome. Identified proteins are shown on the right side. Size markers (in kDa) are indicated on the left.

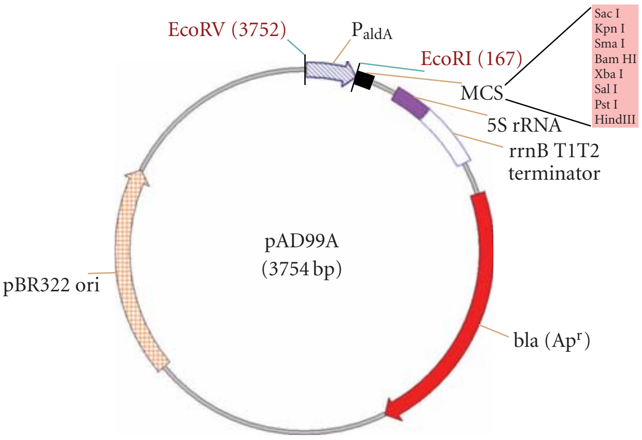


Figure 4: Map of plasmid pAD99A. Its characteristics include pBR322 ori, origin of replication of pBR322; *bla*, ampicillin-resistance gene; $P_{aldA}$, *aldA* promoter; a target gene to be inserted in multiple cloning sites (MCSs).

about 2-fold higher than that obtained by IPTG-inducible *tac* or *udp* promoter. Under the IPTG-inducible promoter, the expression of GFP was low even in LB medium compared to the oleic acid-inducible *aldA* promoter (data not shown). Additionally, the final cell concentration of recombinant *E. coli* W3110 cells in the presence of oleic acid was 4-fold higher than that of recombinant *E. coli* cultured in glucose as a carbon source. This result was further confirmed by fluorescence intensity measurements and confocal microscopy (see Figure 6). Strong fluorescence was uniformly detected in recombinant *E. coli* cells under the control of *aldA* promoter. The fluorescence intensity of GFP obtained from W3110 harboring pAD99GFP was more than 30-fold higher than that obtained from W3110 harboring pTac99GFP, indicating that the *aldA* promoter efficiently enhances recombinant protein production compared to the *tac* promoter. Since the *aldA* promoter was not activated by glucose, GFP was not produced in the presence of glucose in accordance with the proteome profiles. GFP was only produced under the *aldA* promoter along with the supply of exogenous oleic acid. However, the *tac* promoter was not tightly controlled, leaking the recombinant protein even without IPTG induction. These results manifest that the *aldA* promoter is very efficient for the production of recombinant proteins in *E. coli* as an inducible promoter. The maximum productivity can be achieved when the growth and production phases are separated as conducted in this study. Separation of the two phases is often achieved by delaying induction time until the cell density reaches a suitable value.

Until now, the IPTG-inducible promoters *tac* or *trc* have been widely used for basic research. However, the use of IPTG for the large-scale production of human therapeutic proteins is undesirable because of its toxicity and relatively high cost; the high concentration of IPTG can inhibit cell growth and recombinant protein production [22, 23]. Therefore, determination of optimal induction time point as well as the inducer concentration is crucial to increase the overall productivity of recombinant protein. In this regard, there have been significant efforts to overcome such problems by using different inducible or even constitutive expression systems [24]. Thus, the *aldA* promoter found in this study can be effectively used for the enhanced production of recombinant proteins with the aforementioned problems largely resolved.

In summary, the *aldA* promoter satisfies requirements for its utilization as a promoter. First, it is tightly controllable with an appropriate inducer, oleic acid in this case. Tight regulation of the promoter is essential for the synthesis of proteins which may be detrimental to the host cell. For example, the toxic rotavirus VP7 protein effectively kills cells, and must be produced under tightly regulated conditions [25]. Second, its expression is strong and long lasting, resulting in the accumulation of the target protein constituting up to 50% of the total cellular proteins. The third important characteristic of *aldA* promoter is its inducibility in a cost-effective manner by using exogenous oleic acid as an inducer.

## 4. CONCLUSION

Proteome analysis of the cells with focus on proteins induced or repressed by the stimulus provides clues to the understanding of cellular responses. This study revealed that 52 proteins showed significantly altered levels in *E. coli* grown with oleic acid compared to the glucose. Based on the resulting proteome profiles, the promoter of *aldA* gene was

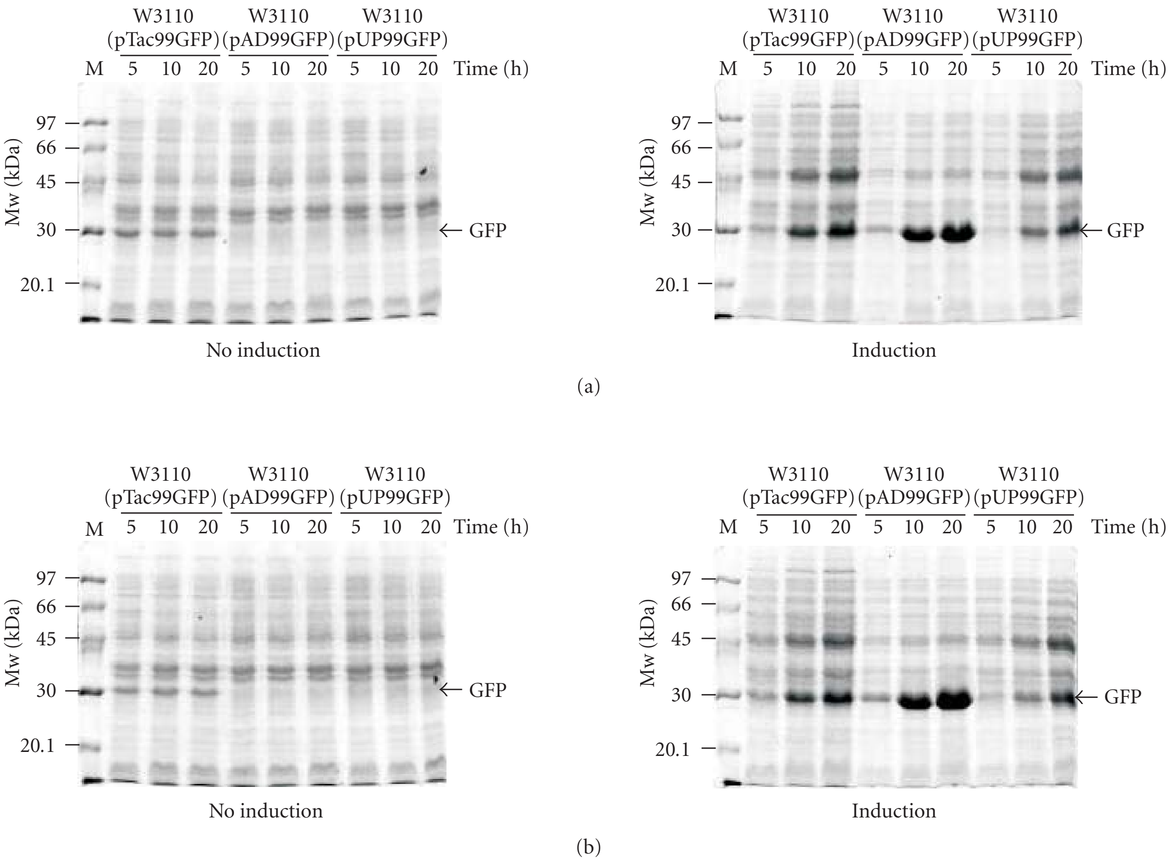


Figure 5: The effect of the recombinant protein production by oleic acid-inducible promoter in recombinant *E. coli* W3110. The cells are induced by exchanging medium supplemented with glucose into the one with oleic acid at the $OD_{600}$ of 0.7 (a) or 1.2 (b). For the control, cells harboring the plasmid containing *tac* promoter were added with 1 mM IPTG at the same values of $OD_{600}$ in the defined medium supplemented with glucose. After induction by IPTG or oleic acid, cells were further cultured for 5, 10, 20 hours, and harvested for 12% (w/v) SDS-PAGE. The arrows indicate the green fluorescent protein (GFP; 26.9 kDa). Size markers (in kDa) are also indicated.

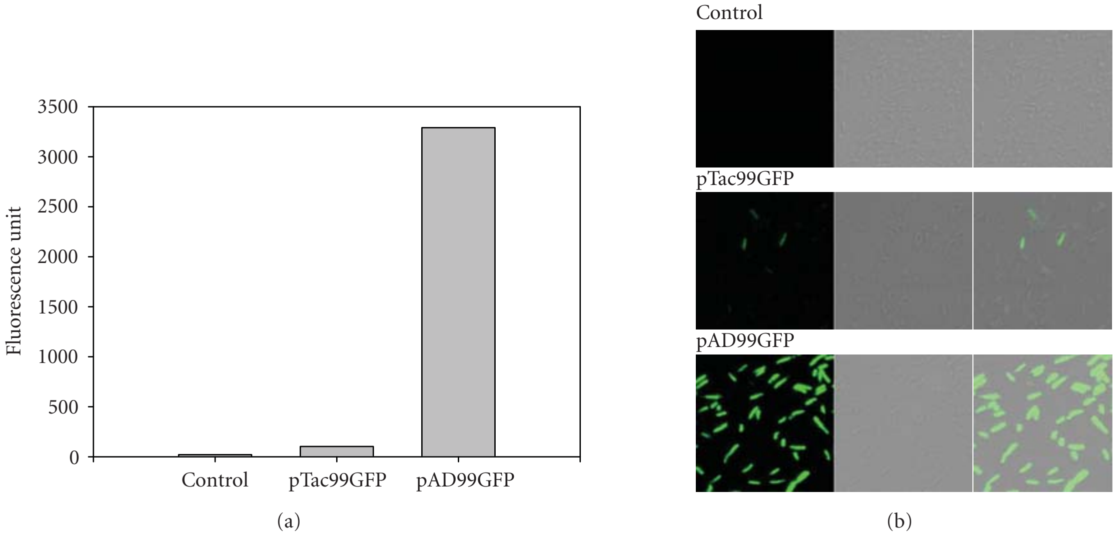


Figure 6: The fluorescence intensities (a) and confocal microscopic images (b) of *E. coli* W3110 cells by induction with IPTG (middle panels) or oleic acid (bottom panels). As a control, the *E. coli* strain without plasmid is also shown (top panels). Shown in (b) are immunofluorescence micrographs (left panels), differential interference micrographs (middle panels), and merged images (right panels) of wild-type *E. coli* W3110 and its recombinant cells harboring pTac99GFP and pAD99GFP.

found to be strongly activated by oleic acid and subsequently to be demonstrated useful as an inducible promoter for the enhanced production of desirable targets. Thus, this study demonstrates that *E. coli* proteome profiles not only provide invaluable information for physiological status of the organism under specific conditions but also propose its biotechnological applications.

## ACKNOWLEDGMENTS

We thank Z. W. Lee (Korea Basic Science Institute, Daejeon, Republic of Korea) for his help with the confocal microscopy. This work was supported by the Korean Systems Biology Research Grant (M10309020000-03B5002-00000) of the Ministry of Science and Technology. Further supports by LG Chem Chair Professorship, IBM SUR program, Microsoft, and by the KOSEF through the Center for Ultramicrochemical Process Systems are appreciated.

## REFERENCES

[1] P. N. Black and C. C. DiRusso, "Molecular and biochemical analyses of fatty acid transport, metabolism, and gene regulation in *Escherichia coli*," *Biochimica et Biophysica Acta*, vol. 1210, no. 2, pp. 123–145, 1994.

[2] C. C. DiRusso, P. N. Black, and J. D. Weimar, "Molecular inroads into the regulation and metabolism of fatty acids, lessons from bacteria," *Progress in Lipid Research*, vol. 38, no. 2, pp. 129–197, 1999.

[3] D. P. Clark and J. E. Cronan Jr., "Two-carbon compounds and fatty acids as carbon sources," in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, et al., Eds., pp. 343–357, ASM Press, Washington, DC, USA, 1996.

[4] C. C. DiRusso, A. K. Metzger, and T. L. Heimert, "Regulation of transcription of genes required for fatty acid transport and unsaturated fatty acid biosynthesis in *Escherichia coli* by FadR," *Molecular Microbiology*, vol. 7, no. 2, pp. 311–322, 1993.

[5] J. W. Campbell and J. E. Cronan Jr., "*Escherichia coli* FadR positively regulates transcription of the *fabB* fatty acid biosynthetic gene," *Journal of Bacteriology*, vol. 183, no. 20, pp. 5982–5990, 2001.

[6] N. Raman, P. N. Black, and C. C. DiRusso, "Characterization of the fatty acid-responsive transcription factor FadR. Biochemical and genetic analyses of the native conformation and functional domains," *Journal of Biological Chemistry*, vol. 272, no. 49, pp. 30645–30650, 1997.

[7] M.-J. Han and S. Y. Lee, "The *Escherichia coli* proteome: past, present, and future prospects," *Microbiology and Molecular Biology Reviews*, vol. 70, no. 2, pp. 362–439, 2006.

[8] A. Matin, "Starvation promoters of *Escherichia coli*. Their function, regulation, and use in bioprocessing and bioremediation," *Annals of the New York Academy of Sciences*, vol. 721, pp. 277–291, 1994.

[9] S. J. Park and S. Y. Lee, "Identification and characterization of a new enoyl coenzyme a hydratase involved in biosynthesis of medium-chain-length polyhydroxyalkanoates in recombinant *Escherichia coli*," *Journal of Bacteriology*, vol. 185, no. 18, pp. 5391–5397, 2003.

[10] S. C. Makrides, "Strategies for achieving high-level expression of genes in *Escherichia coli*," *Microbiological Reviews*, vol. 60, no. 3, pp. 512–538, 1996.

[11] J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 1989.

[12] U. K. Laemmli, "Cleavage of structural proteins during the assembly of the head of bacteriophage T4," *Nature*, vol. 227, no. 259, pp. 680–685, 1970.

[13] M.-J. Han, J. W. Lee, and S. Y. Lee, "Enhanced proteome profiling by inhibiting proteolysis with small heat shock proteins," *Journal of Proteome Research*, vol. 4, no. 6, pp. 2429–2434, 2005.

[14] M. M. Bradford, "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein dye binding," *Analytical Biochemistry*, vol. 72, no. 1-2, pp. 248–254, 1976.

[15] J. W. Lee, S. Y. Lee, H. Song, and J.-S. Yoo, "The proteome of *Mannheimia succiniciproducens*, a capnophilic rumen bacterium," *Proteomics*, vol. 6, no. 12, pp. 3550–3566, 2006.

[16] M.-J. Han, S. S. Yoon, and S. Y. Lee, "Proteome analysis of metabolically engineered *Escherichia coli* producing poly(3-hydroxybutyrate)," *Journal of Bacteriology*, vol. 183, no. 1, pp. 301–308, 2001.

[17] H. Nikaido, "Outer membrane," in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, et al., Eds., pp. 29–47, ASM Press, Washington, DC, USA, 1996.

[18] J. Neuhard and R. A. Kelln, "Biosynthesis and conversions of pyrimidines," in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, et al., Eds., pp. 580–599, ASM Press, Washington, DC, USA, 1996.

[19] E. C. C. Lin, "Dissimilatory pathways for sugars, polyols, and carboxylates," in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, et al., Eds., pp. 307–342, ASM Press, Washington, DC, USA, 1996.

[20] F. X. Quintilla, L. Baldoma, J. Badia, and J. Aguilar, "Aldehyde dehydrogenase induction by glutamate in *Escherichia coli*. Role of 2-oxoglutarate," *European Journal of Biochemistry*, vol. 202, no. 3, pp. 1321–1325, 1991.

[21] S. Y. Lee, "High cell-density culture of *Escherichia coli*," *Trends in Biotechnology*, vol. 14, no. 3, pp. 98–105, 1996.

[22] J. H. Choi, K. J. Jeong, S. C. Kim, and S. Y. Lee, "Efficient secretory production of alkaline phosphatase by high cell density culture of recombinant *Escherichia coli* using the *Bacillus sp.* endoxylanase signal sequence," *Applied Microbiology and Biotechnology*, vol. 53, no. 6, pp. 640–645, 2000.

[23] K. J. Jeong and S. Y. Lee, "High-level production of human leptin by fed-batch cultivation of recombinant *Escherichia coli* and its purification," *Applied and Environmental Microbiology*, vol. 65, no. 7, pp. 3027–3032, 1999.

[24] V. Chauhan, A. Singh, S. M. Waheed, S. Singh, and R. Bhatnagar, "Constitutive expression of protective antigen gene of *Bacillus anthracis* in *Escherichia coil*," *Biochemical and Biophysical Research Communications*, vol. 283, no. 2, pp. 308–315, 2001.

[25] K. R. Emslie, J. M. Miller, M. B. Slade, P. R. Dormitzer, H. B. Greenberg, and K. L. Williams, "Expression of the rotavirus SA11 protein VP7 in the simple eukaryote *Dictyostelium discoideum*," *Journal of Virology*, vol. 69, no. 3, pp. 1747–1754, 1995.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 327468, 7 pages
doi:10.1155/2008/327468

*Research Article*

# Characterization of Phototransduction Gene Knockouts Revealed Important Signaling Networks in the Light-Induced Retinal Degeneration

**Jayalakshmi Krishnan,[1] Gwang Lee,[1,2] Sang-Uk Han,[1,3] and Sangdun Choi[1,4]**

[1] *Department of Molecular Science and Technology, College of Natural Sciences, Ajou University , Suwon 443-749, South Korea*
[2] *Brain Disease Research Center, Ajou University School of Medicine, Suwon 443-749, South Korea*
[3] *Department of Surgery, Ajou University School of Medicine, Suwon 443-749, South Korea*
[4] *Department of Biological Sciences, College of Natural Science, Ajou University, Suwon 443-749, South Korea*

Correspondence should be addressed to Sangdun Choi, sangdunchoi@ajou.ac.kr

Received 24 August 2007; Accepted 17 December 2007

Recommended by Daniel Howard

Understanding the molecular pathways mediating neuronal function in retinas can be greatly facilitated by the identification of genes regulated in the retinas of different mutants under various light conditions. We attempted to conduct a gene chip analysis study on the genes regulated during rhodopsin kinase (Rhok$^{-/-}$) and arrestin (Sag$^{-/-}$) knockout and double knockouts in mice retina. Hence, mice were exposed to constant illumination of 450 lux or 6,000 lux on dilated pupils for indicated periods. The retinas were removed after the exposure and processed for microarray analysis. Double knockout was associated with immense changes in gene expression regulating a number of apoptosis inducing transcription factors. Subsequently, network analysis revealed that during early exposure the transcription factors, p53, c-MYC, c-FOS, JUN, and, in late phase, NF-$\kappa$B, appeared to be essential for the initiation of light-induced retinal rod loss, and some other classical pro- and antipoptotic genes appeared to be significantly important as well.

## 1. INTRODUCTION

The molecular analysis of knockouts provides us with a plenty of knowledge on the functions of genes in mammals. Thus, the characterization of knockouts in mouse retinas is of great importance in our understanding of the mechanisms of signaling networks in the visual system. Rods and cones in vertebrate retina transform visual information into neuronal signals. In mouse rod photoreceptors, light activates rhodopsin, a G-protein-coupled receptor, which is then phosphorylated by rhodopsin kinase [1–4]. Visual arrestin terminates the light response by selectively binding to phosphorylated rhodopsin [5, 6]. Upon illumination and transducin, a G-protein specific to rod photoreceptor cells turns on and calcium influx occurs [7].

Alternatively in mammals, exposure to light can induce photoreceptor cell death and retinal degeneration. The retina of transgenic mice with a null mutation in the gene encoding rhodopsin kinase [8] or arrestin [9] had been sensitized to light damage [10] and revealed prolonged rhodopsin signaling. Furthermore, mouse rod photoreceptor cells lacking the $\alpha$-subunit of transducin revealed that light-activated rhodopsin and phototransduction signaling were no longer connected [11]. In addition, under certain conditions, the absence of c-FOS [12] or the absence [13] or modification [14] of Rpe65 prevented light-induced degeneration. In previous studies, two different pathways of photoreceptor-cell apoptosis induced by light, transducin-dependent (low light), and AP-1 dependent (bright light), were suggested [15]. Excessive levels of light induced caspase-independent photoreceptor apoptosis have also been proposed during retinal development [16]. However, the molecular signaling networks that initiate the retinal degeneration cascade are not fully understood [17, 18].

The rationale of the study was to delineate the signal transduction networks by taking account of the gene expression changes at different time points and light intensities. In this study, two key gene knockouts in phototransduction, such as rhodopsin kinase (Rhok$^{-/-}$), arrestin (Sag$^{-/-}$), and rhodopsin kinase/arrestin (Rhok$^{-/-}$/Sag$^{-/-}$), were tested by measuring the expression levels of thousands of genes for their roles in phototransduction signaling in light-induced retinal degeneration.

## 2. MATERIALS AND METHODS

### 2.1. Animals

All procedures concerning animals were performed in accordance with the Association for Research in Vision and Ophthalmology (ARVO, MD, USA) Statement on the use of animals in ophthalmic and vision research. Rhodopsin kinase (Rhok$^{-/-}$) and arrestin (Sag$^{-/-}$) knockout mice were generated [8, 9]. These mice were crossed to each other to obtain the double-deficient mice, rhodopsin kinase arrestin (Rhok$^{-/-}$/Sag$^{-/-}$). All mice including wild-type (WT) were reared in dark until the given experiments were performed. Wild-type mice were derived from an initial cross of 129Sv and C57BL/6. The mice used in this study ranged from 6 to 8 weeks of age.

### 2.2. Light illumination

The mice reared in dark were placed in aluminum foil-wrapped polycarbonate cages that were covered with stainless steel wire tops to protect them from uncontrolled light exposure. Fluorescent lamps gave off light from an opening at the top of the cage. They were supplied with food and water at the bottom of the cage. Constant illumination of 450 lux on dilated pupils (1% Cyclogyl, Alcon; 5% Phenylephrine, Ciba Vision), or 6 000 lux on dilated pupils for indicated periods (1 hour for 450 and 80 minutes for 6 000 lux) was generated by diffuse, cool, white florescent lamps. The temperature was kept at 25°C during irradiation. After light exposure, the mice retinas were either analyzed immediately or after a given period in darkness. Retinas were removed rapidly through a slit in the cornea and frozen in liquid nitrogen until total RNA was extracted by the Trizol method (Invitrogen Life Technologies). The retinas from three to four mice were pooled to make the corresponding sample.

### 2.3. Microarray analysis

With 3 $\mu$g of total RNA from retinas as starting material, first strand cDNA was synthesized using T7-oligo dT primer and SuperScript II (Invitrogen Life Technologies). Second strand cDNA was synthesized with second strand buffer (Invitrogen Life Technologies), DNA polymerase I (New England Biolabs), DNA ligase (New England Biolabs), and RNase H (Invitrogen Life Technologies). cDNA was extracted using phenol:chloroform:isoamyl alcohol, precipitated with ethanol, washed with 80% and 100% cold ethanol, and air dried. The dried pellet was then dissolved in 22 $\mu$L of nuclease-free water and stored at −20°C. In vitro transcription was performed using the RNA Transcript Labeling Kit (Enzo Diagnostics) to produce hybridizable biotin-labeled RNA targets. The cDNA was used as a template in the presence of a mixture of unlabeled NTPs and biotinylated CTP and UTP. After in vitro transcription, cRNA was purified using RNeasy Mini Kit (Qiagen Inc.). The fragmented cRNA, generated by incubating at 94°C for 35 minutes, was applied to the Affymetrix GeneChip U74Av2 array (total 12,488 probe sets) and hybridized at 40°C for 16 hours. After hybridization, the array was washed several times and stained with streptavidine-conjugated phycoerythrin in the GeneChip Fluidics Station 400 (Affymetrix, Inc.). The arrays were scanned by the Agilent Scanner (Agilent Technologies) and analyzed with GeneChip Analysis Suite 5.0 (Affymetrix, Inc.).

### 2.4. Network analysis

For each array, genes that were regulated more than or equal to 0.5 and less than or equal to −0.5 in $\log_2$ ratio were loaded onto the Ingenuity Pathways Analysis program (http://analysis.ingenuity.com) to identify possible gene networks or pathways.

## 3. RESULTS AND DISCUSSION

### 3.1. The general patterns of regulation

To identify the genes regulated by different knockout conditions, we performed DNA microarrays. The number of changed genes (≥2 folds) in each condition was measured (see Figure 1). The initial screening was done in the c-Fos$^{-/-}$ knockout as well as the transducin (Gnat$^{-/-}$) knockout to understand the nature of disturbance in signaling mechanisms (manuscript in preparation). However, herewith we will be discussing on the initial changes in gene expression with special focus on Rhok$^{-/-}$/Sag$^{-/-}$ knockout mice. The experiments were also extended to include a low light condition (450 lux) and a high light condition (6 000 lux) for a series of dark adaptation periods up to 24 hours in wildtype and Rhok$^{-/-}$/Sag$^{-/-}$ mice, which might provide a visional molecular progress on the mechanisms of light-induced apoptosis condition.

Wildtype and knockout (Rhok$^{-/-}$/Sag$^{-/-}$) mice were exposed to low light (450 lux) with dilated pupils for 1 hour and again placed back into a dark room for up to 20 hours (see Figure 1). A subgroup of wildtype mice were also exposed to bright light (6 000 lux) with dilated pupils for 80 minutes and placed back into a dark room for an indicated period of time (up to 24 hours). There were only a few elements (tens out of total 12 488 probe sets) regulated in wildtype under low light (450 lux) condition during the test up to 20 hours. However, mutants (Rhok$^{-/-}$/Sag$^{-/-}$) placed under low light (450 lux) for 1 hour and wildtype mice placed under bright light (6 000 lux) for 80 minutes had considerably more changed genes with the lapse of time. One notion is that only mutant (Rhok$^{-/-}$/Sag$^{-/-}$) can cause apoptosis in this condition and there were significantly different molecular changes noted in mutants, which were different from wildtype.

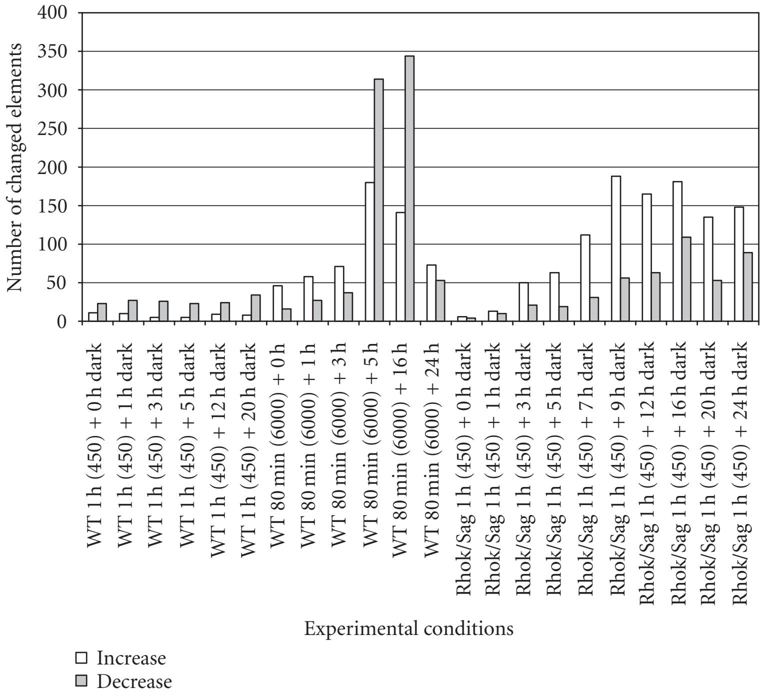


Figure 1: Number of upregulated and downregulated genes in wild-type and rhodopsin kinase/arrestin knock-out (Rhok$^{-/-}$/Sag$^{-/-}$) after dark adaptation for indicated periods.

### 3.2. *Significant early phase gene expression in Rhok$^{-/-}$/Sag$^{-/-}$ mutant*

As shown in Table 1, many transcription factors were highly upregulated at the early stage of dark adaptation in the Rhok$^{-/-}$/Sag$^{-/-}$ mutant exposed to low light (450 lux with dilation) for 1 hour. For example, FBJ osteosarcoma oncogene (c-FOS), CCAAT/enhancer binding protein delta (C/EBP delta), early growth response 1 (EGR1), brain derived neurotrophic factor (BDNF), activating transcription factor 4 (ATF4), fos-like antigen 1 (FRA1), activating transcription factor 3 (ATF3), and growth arrest and DNA-damage-inducible 45 beta (GADD45 beta) were highly expressed. The induction of these transcription factors could potentially trigger the production of their downstream target genes. These genes may have been regulated by different mechanisms, yet coordinately expressed at an early stage in the light-exposed mutant mice (Rhok$^{-/-}$/Sag$^{-/-}$) when retinal degeneration occurs and, therefore, may have roles to cooperatively play in light-induced apoptotic cellular signaling networks.

### 3.3. *The transcriptional regulations in the apoptosis pathway surveyed by GenMAPP*

Studying the signal transduction pathways and transcriptional regulation using DNA microarray can be a tremendous challenge to biologists. Obviously, novel bioinformatic tools are required to gain biological insights out of microarray data. Here, we surveyed the transcriptional regulation of the apoptosis pathway focusing on Rhok$^{-/-}$/Sag$^{-/-}$ mice using the GenMAPP program (see Figure 2) [19]. In wild-types, including both low and bright light conditions, none of the genes in the hypothetical apoptosis pathway provided by the GenMAPP were differentially expressed. However, we observed a serial regulation of genes involved in apoptosis in Rhok$^{-/-}$/Sag$^{-/-}$ mutants. After 1 hour of light exposure (450 lux with dilation) followed by 3 hours of dark adaptation, some AP1 components including c-JUN were highly upregulated. This upregulation was observed till 5 hours of dark adaptation, disappeared at 7 hours, and never surfaced again till 24 hours, which was the whole duration of the experiment. Instead, I$\kappa$B and NF-$\kappa$B p105 subunit were upregulated and these high expression levels were maintained for a specific period of time. I$\kappa$B expression was maintained for 24 hours while NF-$\kappa$B p105 expression went back to normal at 20 hours of dark adaptation. AP1 was reported to be an essential component in light-induced apoptosis [12, 15], which seems to be authentic in the Rhok$^{-/-}$/Sag$^{-/-}$ mouse model system. The expression level of another AP1 component, c-FOS, also turned out to be highly induced after 1 hour of light exposure and maintained for at least 24 hours in Rhok$^{-/-}$/Sag$^{-/-}$ mice (see Table 1). This c-FOS was also induced in wildtype mice under the both low and high light conditions and further studies in our lab are delineating this process.

### 3.4. *The molecular functions of regulated genes in light-induced apoptosis*

Cluster Assignment for Biological Inference (CLASSIFI) analysis [20] was performed on wild-type (450 lux and 6 000 lux) and mutant (Rhok$^{-/-}$/Sag$^{-/-}$) time point data.

Table 1: Significant or specific regulations in early stage of Rhok$^{-/-}$/Sag$^{-/-}$ mutant. L0: low light for 0 hr in wild-type; L1: low light for 1 hr in wild-type; and so on. H0, high light for 0 hr in wild-type; H1: high light for 1 hr in wild-type; and so on. M0: low light for 0 hr in Rhok$^{-/-}$/Sag$^{-/-}$ mutant; M1: low light for 1 hr in Rhok$^{-/-}$/Sag$^{-/-}$ mutant; and so on. The numbers are in $\log_2$ ratio. Refer to the text for details.

| Gene Name | Accession | L0 | L1 | L3 | L5 | L12 | L20 | H0 | H1 | H3 | H5 | H16 | H24 | M0 | M1 | M3 | M5 | M7 | M9 | M12 | M16 | M20 | M24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FBJ osteosarcoma oncogene | V00727 | 2.8 | 1 | 0 | 1.6 | 1.5 | 0 | 2 | 1.5 | 3 | 3.4 | 2.2 | 0.7 | 1.5 | 1.4 | 2.5 | 1.9 | 2.5 | 2.6 | 2.1 | 2.1 | 1.7 | 1.1 |
| connective tissue growth factor | M70642 | 0.9 | 0 | 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 1.4 | 0 | −1.3 | 1.1 | 1.8 | 1.1 | 1 | 1.9 | 1.7 | 1.6 | 1.9 | 1.3 | 1.3 |
| CCAAT/enhancer binding protein (C/EBP), delta | X61800 | 0 | 0 | 0 | 0 | −0.6 | −0.5 | 0 | 0 | 2.6 | 3.6 | 3 | 2.9 | 1.1 | 0 | 2.5 | 2.9 | 3.6 | 4.2 | 4 | 4.1 | 4 | 4.3 |
| early growth response 1 | M28845 | 2.5 | 0 | 1.1 | 1.8 | 1.5 | 0 | 0 | 2.3 | 3 | 3.2 | 1.6 | 0 | 1 | 1.4 | 2.8 | 1.5 | 2.4 | 2.9 | 2.4 | 2.4 | 1.4 | 0.6 |
| prostaglandin D2 synthase (21 kDa, brain) | AB006361 | −0.7 | −0.8 | 0 | −0.8 | −0.5 | −0.3 | 0 | 1 | 1.2 | 0.9 | 0 | −0.3 | 0.9 | 1 | 1.2 | 1 | 0.9 | 0.5 | 0.9 | 1.1 | 0.8 | 0.6 |
| brain derived neurotrophic factor | X55573 | 0.6 | 0 | 0 | 0 | 0 | 0 | 1 | 0.8 | 0 | 1.3 | 1.4 | 0.6 | 0.9 | 0 | 0 | 0 | 1 | 1.2 | 0.8 | 1.2 | 0.6 | 0.9 |
| nuclear receptor subfamily 4, group A, member 1 | X16995 | 2.1 | 0.7 | 0 | 0.4 | 0.7 | 0 | 1.3 | 0 | 0 | 1.2 | 0 | 0 | 0.9 | 0 | 0 | 0 | 0 | 0.4 | 0.5 | 0.6 | 0.9 | 0 |
| prostaglandin D2 synthase (21 kDa, brain) | AB006361 | −0.5 | −1 | 0.4 | −0.8 | −0.8 | −0.4 | 0.6 | 1 | 1.3 | 1.1 | 0 | 0 | 0.8 | 1.2 | 1.2 | 0.5 | 0.7 | 0.6 | 1 | 1.1 | 0.9 | 0.8 |
| activating transcription factor 4 | M94087 | 1 | 1 | 0.5 | 0 | 0 | 0 | 0.7 | 0.7 | 0 | 0 | 0 | 0 | 0.7 | 1.2 | 1.3 | 1.2 | 1.1 | 0.6 | 0.6 | 0.5 | 0 | 0 |
| retinol binding protein 1, cellular | X60367 | 0 | −0.4 | 0 | −0.4 | 0 | 0 | 0.6 | 0.9 | 0.6 | 0.9 | 1 | 0.6 | 0.7 | 0.6 | 0.6 | 0 | 0.9 | 1.3 | 1.8 | 2.2 | 1.9 | 2.4 |
| metallothionein 2 | K02236 | 0 | 0 | 0.9 | 0 | 0 | 0 | 0 | 1.4 | 2 | 3.7 | 0 | 1.2 | 0.6 | 1.2 | 2.5 | 2.6 | 4 | 3.8 | 4.1 | 3.7 | 3.1 | 2.7 |
| metallothionein 1 | V00835 | 0 | 0.6 | 1 | 0.4 | −0.4 | 0 | 0 | 0.9 | 1.7 | 3 | 1 | 0.8 | 0 | 1.1 | 2.1 | 2.5 | 3.7 | 3.8 | 3.7 | 4 | 2.6 | 2.6 |
| Bcl2-associated athanogene 3 | AV373612 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.7 | 2.8 | 0 | 0 | 0 | 0 | 5.2 | 0 | 6.4 | 6.6 | 6 | 6.8 | 6.1 | 6.2 |
| fos-like antigen 1 | AF017128 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3.2 | 0 | 0 | 0 | 0 | 3.7 | 0 | 4.6 | 4.8 | 4.9 | 4 | 3.7 | 0 |
| growth arrest and DNA-damage-inducible 45 beta | AV138783 | 1.2 | 0.9 | 0 | 0 | 0 | 0 | 1.3 | 1.1 | 2.2 | 3.3 | 3 | 2 | 0 | 0 | 3.1 | 2 | 3.6 | 4.4 | 4.5 | 4.2 | 3.7 | 3.5 |
| activating transcription factor 3 | U19118 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2.9 | 2.5 | 2.3 | 2.6 | 2.7 | 2.9 | 2.7 |
| cytokine inducible SH2-containing protein 3 | AV374868 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.8 | 3.1 | 2.9 | 2.9 | 0 | 0 | 2.6 | 2.4 | 3.3 | 3.2 | 3.7 | 3.3 | 3.3 | 3.3 |
| myocyte enhancer factor 2C | AI426400 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.6 | −1 | −0.9 | 0 | 0 | −0.7 | −1.2 | 0 | −0.5 | −0.7 | −0.7 | −1.3 | −1.3 | −1.7 |
| Similar to rhodopsin (opsin 2, rod pigment) | M36699 | 0 | −0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.3 | 0 | −1.2 | −1.4 | −1.6 | −1.2 | −1.4 | −1.5 | −0.9 | −1.5 |
| hexokinase 2 | Y11666 | 0 | 0 | 0 | 0 | 0 | 0 | −0.5 | −0.7 | 0 | 0 | 0 | −0.7 | −0.4 | −1 | −1.3 | −1.2 | −1.3 | −1 | −1.2 | −1 | −0.6 | −1.2 |
| expressed sequence AA960287 | AW061237 | −0.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.6 | 0 | 0 | −0.6 | 0 | −0.9 | −1 | −1.4 | −0.8 | −1.1 | −1.3 | −0.8 | −1.2 |
| rod outer segment membrane protein 1 | AV356715 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.4 | 0 | 0 | 0 | −0.3 | −0.6 | −0.6 | −1 | −0.9 | −1.5 | −1 | −1 | −1 | −0.6 | −0.9 |
| WNT1 inducible signaling pathway protein 1 | AF100777 | −0.4 | 0 | 0 | 0 | 0 | 0 | −0.3 | −0.5 | −0.6 | 0 | 0 | 0 | −0.6 | −0.9 | −0.8 | 0 | −0.9 | −1.1 | −1.4 | −1.4 | −1 | −1.5 |
| high mobility group box 2 | X67668 | 0 | 0 | −0.8 | 0 | 0 | 0 | 0 | −0.7 | −1.2 | 0 | 0 | 0 | −0.6 | −0.9 | −1.3 | −1 | 0 | 0 | −1 | −1 | −0.8 | 0 |
| ESTs | C78037 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.8 | 0 | 0 | 0 | −1.2 | 0 | −0.9 | −1.1 | −0.7 | −1.3 |
| kallikrein 9 | M17979 | 0.6 | 0 | 0 | −0.4 | 0 | −1.4 | 0 | −0.8 | −1.7 | −3.9 | 0 | −3.8 | −0.8 | 0 | −0.9 | −0.8 | −0.7 | 0 | −0.9 | 0 | −1.4 | −0.9 |
| ISL1 transcription factor, LIM/homeodomain, (islet-1) | AJ132765 | −0.2 | −0.5 | 0 | 0 | −0.3 | 0 | −0.5 | −0.7 | 0 | 0 | 0 | 0 | −0.8 | −1.5 | −0.6 | −0.5 | −0.7 | 0 | 0 | 0 | −0.7 | 0 |
| RIKEN cDNA 1110013B16 gene | AW123271 | 0 | 0 | −0.5 | 0 | 0 | −0.3 | 0 | 0 | 0 | 0 | −1.6 | 0 | −1.3 | −0.9 | −1.1 | −1.3 | −1.4 | −0.6 | −1.2 | −1.3 | −1.4 | −1.1 |

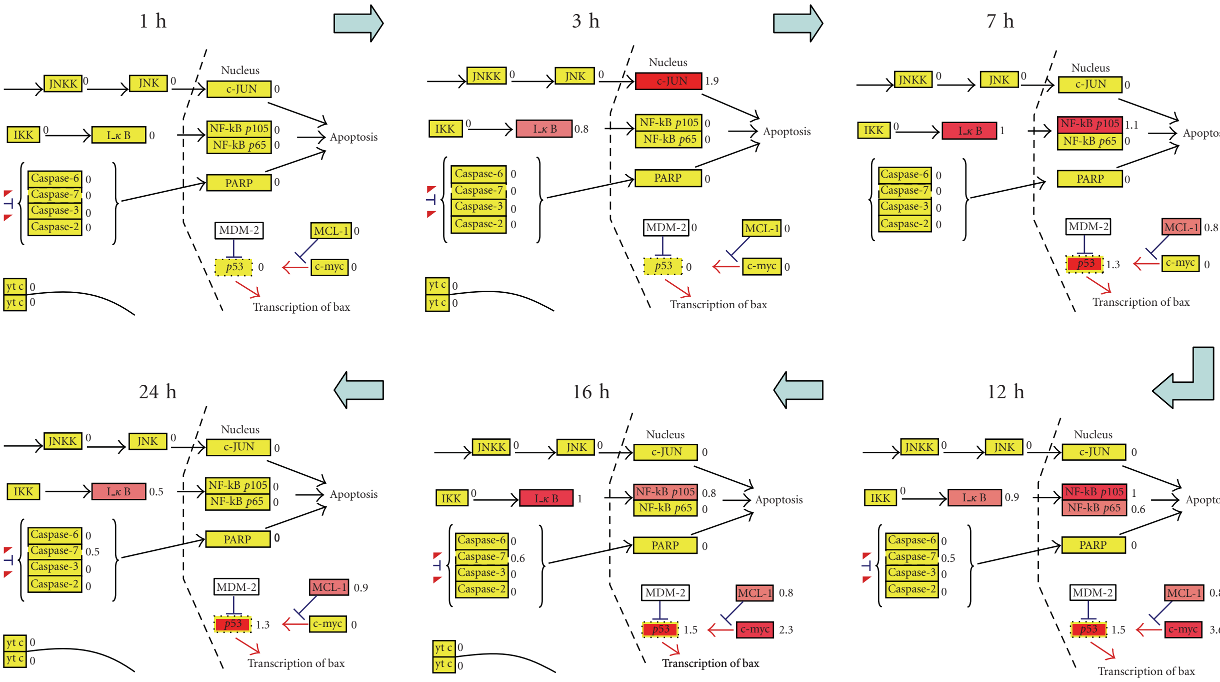


Figure 2: Apoptosis map. Expression data of Rhok$^{-/-}$/Sag$^{-/-}$ mice was overlaid onto a hypothetical apoptosis map using the GenMAPP (http://www.genmapp.org) program. Each rectangle represents a gene in the pathway. The number on the right of the rectangle corresponds to the log $_2$ ratio. Genes highlighted in yellow are unchanged. Pink and red denote upregulated genes while dark and light green correspond to downregulated genes. Rhok$^{-/-}$/Sag$^{-/-}$ mice were exposed to 1 hour of low light (450 lux) and adapted for the indicated time period of darkness.

There were not any significant GO clusters for the wild-type mice but the case is opposite in the mutant data with many significant GO coclustering in Figure 3. There were many defense/immune response-related gene clusters (highly upregulated genes). There is also one vision-related cluster with mostly downregulated genes, indicating photoreceptor cell damages. The significant cutoff use for this CLASSIFI analysis was $2.17\times10^{-5}$, estimated using the Bonnferoni correction with an alpha of 0.05.

### *3.5. The expression of other classical pro- and antiapoptotic genes*

Surprisingly, we have found the expressions of other classical pro- and antiapoptotic genes, which were different from what we might have expected. For example, the expression levels of proapoptotic genes, caspase 1, 2, 3, 6, 7, 8, 9, 11, 12, and 14 did not change at all in Rhok$^{-/-}$/Sag$^{-/-}$ mice when retina cells degeneration occurred after light exposure. The expression levels of these caspases did not change in wild-type under both low (450 lux) and bright (6 000 lux) light conditions. Previously, it has been reported that the expression of caspase 3 was distinctly upregulated in blue light-induced apoptosis in photoreceptor cells [21].

There were other elements such as Bcl2 families which were distinctly working on cell death. Our data indicate that Bcl2 family including BAX, BCL2L10, BAD, BAK1, BAG3, BOK, BAL2L, BCL2L11, BAG1, BCL2L2, and BAD were found to be unchanged, suggesting that the light-induced enzymatic apoptosis may not be regulated at the level of transcription but rather by the activities of proteins that were already present normally. On the other hand, there was also the generation of new transcripts of active molecules which then induced the apoptotic death cascade. p53 and c-MYC were upregulated at 7 or 9 hours after dark adaptation and were maintained at higher levels up to later time points (20 to 24 hours). The transcription factor AP1, c-FOS, and JUN family seemed to be essential for the initiation of light-induced retinal rod loss, while other classical pro- and antiapoptotic genes appeared to be also important in our model system (see Figure 2 and Table 1).

The current study describes about the light-induced gene regulation and transcriptional responses in mouse retinas after genes knockout. As shown in Table 1, many transcription factors were highly upregulated at the early stage of dark adaptation in the Rhok$^{-/-}$/Sag$^{-/-}$ mutant exposed to low light (450 lux with dilation). The induction of transcription factors, such as c-FOS, C/EBP delta, EGR1, BDNF, ATF4, FRA1, ATF3, and GADD45 beta could potentially trigger the production of their downstream target genes. When they were coordinately expressed at early stages in light-exposed mutant mice (Rhok$^{-/-}$/Sag$^{-/-}$), it is likely that they function cooperatively in the cellular signaling networks to induce retinal degeneration. Inferring signal transduction pathways and transcription regulation using DNA microarray data and legendary literature-based interaction information can be a tremendous challenge. While the transcription regulation in the apoptosis pathway derived from the GenMAPP program

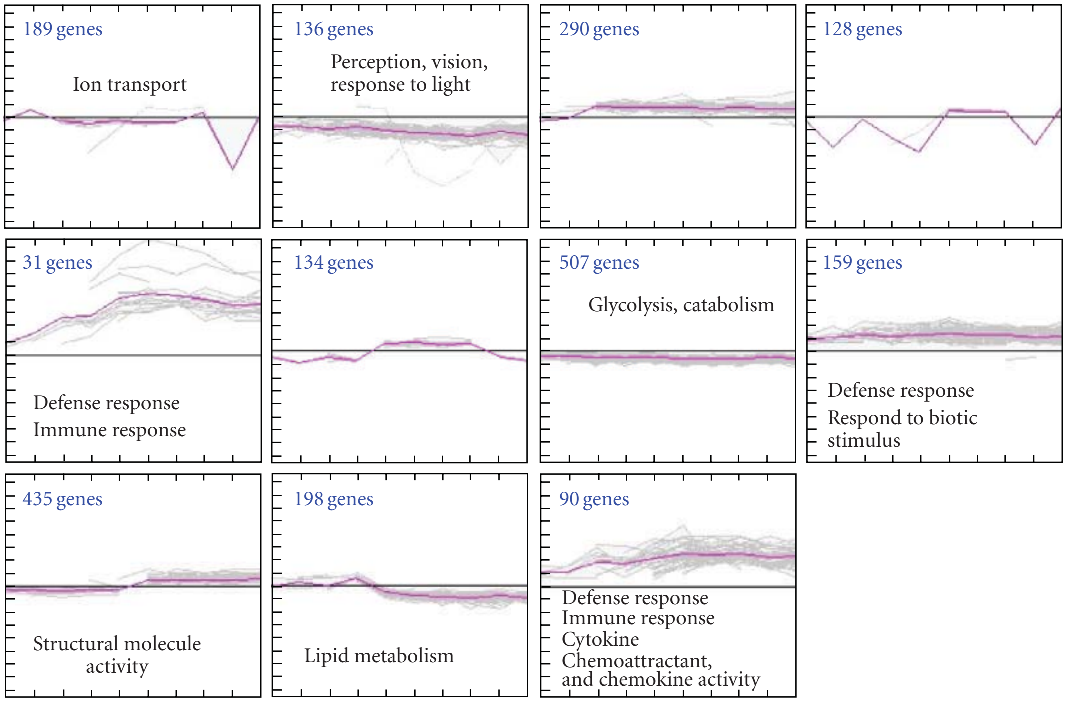


Figure 3: CLASSIFI analysis was performed on mutant (Rhok$^{-/-}$/Sag$^{-/-}$). The significant cutoff use for this CLASSIFI analysis was $2.17 \times 10^{-5}$, estimated using the Bonnferoni correction with an alpha of 0.05.

(see Figure 2) is useful to understand the mechanism, there are many possible networks which we do not know yet. The initial trigger for condition-specific transcription in complex animals often comes from several groups of regulatory transcription factors. According to the dictates of what best contributes to organism survival and selection, individual genes have binding sites to accommodate an assortment of different types of transcription factors. There is no surprise, when complicated biological events are affected, that individual genes are involved in multiple pathways. This signal flow leads to the expression of genes responsible for transcriptional regulation, transport, defense response, immune response, signal transduction, and vision in retinas of Sag$^{-/-}$ mice (see Figure 3).

## 4. CONCLUSION

Mild or excessive light accelerates the cell death process in certain knockout mice. In this study, we found out the signal transduction networks by analyzing the gene expression changes during different time points of key phototransduction gene knockout in mice. Herewith, we revealed many gene transcripts essential for the initiation of light-induced rod degeneration and proposed important networks fabricated in pro- and antiapoptotic signaling.

## ACKNOWLEDGMENT

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-311-C00482) and the Korea Science and Engineering Foundation (KOSEF) Grant, funded by the Korea government (MOST) (R01-2007-000-20533-0) to “S. Choi.” J. Krishnan was supported by BK21 program.

## REFERENCES

[1] V. Y. Arshavsky, “Rhodopsin phosphorylation: from terminating single photon responses to photoreceptor dark adaptation,” *Trends in Neurosciences*, vol. 25, no. 3, pp. 124–126, 2002.

[2] T. Maeda, Y. Imanishi, and K. Palczewski, “Rhodopsin phosphorylation: 30 years later,” *Progress in Retinal and Eye Research*, vol. 22, no. 4, pp. 417–434, 2003.

[3] I. I. Senin, K.-W. Koch, M. Akhtar, and P. P. Philippov, “Ca2+-dependent control of rhodopsin phosphorylation: recoverin and rhodopsin kinase,” *Advances in Experimental Medicine and Biology*, vol. 514, pp. 69–99, 2002.

[4] K. D. Ridge, N. G. Abdulaev, M. Sousa, and K. Palczewski, “Phototransduction: crystal clear,” *Trends in Biochemical Sciences*, vol. 28, no. 9, pp. 479–487, 2003.

[5] P. J. Dolph, “Arrestin: roles in the life and death of retinal neurons,” *Neuroscientist*, vol. 8, no. 4, pp. 347–355, 2002.

[6] W. E. Miller and R. J. Lefkowitz, “Arrestins as signaling molecules involved in apoptotic pathways: a real eye opener,” *Science’s STKE: Signal Transduction Knowledge Environment*, vol. 2001, no. 69, p. pe1, 2001.

[7] C. L. Makino, X.-H. Wen, and J. Lem, “Piecing together the timetable for visual transduction with transgenic animals,” *Current Opinion in Neurobiology*, vol. 13, no. 4, pp. 404–412, 2003.

[8] C.-K. Chen, M. E. Burns, M. Spencer, et al., “Abnormal photoresponses and light-induced apoptosis in rods lacking rhodopsin kinase,” *Proceedings of the National Academy of*

*Sciences of the United States of America*, vol. 96, no. 7, pp. 3718–3722, 1999.

[9] J. Chen, M. I. Simon, M. T. Matthes, D. Yasumura, and M. M. LaVail, "Increased susceptibility to light damage in an arrestin knockout mouse model of Oguchi disease (stationary night blindness)," *Investigative Ophthalmology and Visual Science*, vol. 40, no. 12, pp. 2978–2982, 1999.

[10] S. Choi, W. Hao, C.-K. Chen, and M. I. Simon, "Gene expression profiles of light-induced apoptosis in arrestin/rhodopsin kinase-deficient mouse retinas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 23, pp. 13096–13101, 2001.

[11] P. D. Calvert, N. V. Krasnoperova, A. L. Lyubarsky, et al., "Phototransduction in transgenic mice after targeted deletion of the rod transducin $\alpha$-subunit," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 25, pp. 13913–13918, 2000.

[12] F. Hafezi, J. P. Steinbach, A. Marti, et al., "The absence of $c\text{-}fos$ prevents light-induced apoptotic cell death of photoreceptors in retinal degeneration *in vivo*," *Nature Medicine*, vol. 3, no. 3, pp. 346–349, 1997.

[13] C. Grimm, A. Wenzel, F. Hafezi, S. Yu, T. M. Redmond, and C. E. Remé, "Protection of *Rpe*65-deficient mice identifies rhodopsin as a mediator of light-induced retinal degeneration," *Nature Genetics*, vol. 25, no. 1, pp. 63–66, 2000.

[14] M. Samardzija, A. Wenzel, M. Naash, C. E. Remé, and C. Grimm, "*Rpe*65 as a modifier gene for inherited retinal degeneration," *European Journal of Neuroscience*, vol. 23, no. 4, pp. 1028–1034, 2006.

[15] W. Hao, A. Wenzel, M. S. Obin, et al., "Evidence for two apoptotic pathways in light-induced retinal degeneration," *Nature Genetics*, vol. 32, no. 2, pp. 254–260, 2002.

[16] M. Donovan and T. G. Cotter, "Caspase-independent photoreceptor apoptosis *in vivo* and diffrential expression of apoptotic protease activating factor-1 and caspase-3 during retinal development," *Cell Death & Differentiation*, vol. 9, no. 11, pp. 1220–1231, 2002.

[17] A. Wenzel, C. Grimm, M. Samardzija, and C. E. Remé, "Molecular mechanisms of light-induced photoreceptor apoptosis and neuroprotection for retinal degeneration," *Progress in Retinal and Eye Research*, vol. 24, no. 2, pp. 275–306, 2005.

[18] A. Roca, K.-J. Shin, X. Liu, M. I. Simon, and J. Chen, "Comparative analysis of transcriptional profiles between two apoptotic pathways of light-induced retinal degeneration," *Neuroscience*, vol. 129, no. 3, pp. 779–790, 2004.

[19] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Nature Genetics*, vol. 31, no. 1, pp. 19–20, 2002.

[20] J. A. Lee, R. S. Sinkovits, D. Mock, et al., "Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation," *BMC Bioinformatics*, vol. 7, p. 237, 2006.

[21] J. Wu, A. Gorman, X. Zhou, C. Sandra, and E. Chen, "Involvement of caspase-3 in photoreceptor cell apoptosis induced by *in vivo* blue light exposure," *Investigative Ophthalmology & Visual Science*, vol. 43, no. 10, pp. 3349–3354, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 597913, 30 pages
doi:10.1155/2008/597913

*Review Article*

# Metabolic Control Analysis: A Tool for Designing Strategies to Manipulate Metabolic Pathways

**Rafael Moreno-Sánchez, Emma Saavedra, Sara Rodríguez-Enríquez, and Viridiana Olín-Sandoval**

*Departamento de Bioquímica, Instituto Nacional de Cardiología, Juan Badiano no. 1, Colonia Sección 16, Tlalpan, México DF 14080, Mexico*

Correspondence should be addressed to Rafael Moreno-Sánchez, rafael.moreno@cardiologia.org.mx

Received 1 October 2007; Revised 16 January 2008; Accepted 26 March 2008

Recommended by Daniel Howard

The traditional experimental approaches used for changing the flux or the concentration of a particular metabolite of a metabolic pathway have been mostly based on the inhibition or over-expression of the presumed rate-limiting step. However, the attempts to manipulate a metabolic pathway by following such approach have proved to be unsuccessful. Metabolic Control Analysis (MCA) establishes how to determine, quantitatively, the degree of control that a given enzyme exerts on flux and on the concentration of metabolites, thus substituting the intuitive, qualitative concept of rate limiting step. Moreover, MCA helps to understand (i) the underlying mechanisms by which a given enzyme exerts high or low control and (ii) why the control of the pathway is shared by several pathway enzymes and transporters. By applying MCA it is possible to identify the steps that should be modified to achieve a successful alteration of flux or metabolite concentration in pathways of biotechnological (e.g., large scale metabolite production) or clinical relevance (e.g., drug therapy). The different MCA experimental approaches developed for the determination of the flux-control distribution in several pathways are described. Full understanding of the pathway properties when is working under a variety of conditions can help to attain a successful manipulation of flux and metabolite concentration.

## 1. INTRODUCTION

Is an effort to manipulate the metabolism of an organism worthy and reasonable, knowing that this cellular process has been continuously modified and refined through evolution and natural selection for adapting, in the most convenient manner, to the ongoing environmental conditions? The answer to this question seems obvious when three broad areas of research and development are identified in which manipulation of metabolic pathways is relevant: (a) drug design to treat diseases, (b) genetic engineering of organisms of biotechnological interest, and (c) genetic syndromes therapy.

Historically, drug design was the first area in which modification of metabolism was tried: the primary goal of drug administration is the inhibition of essential metabolic pathways, for example, in a parasite or a tumor cell. Thus, any metabolic pathway can be a potential therapeutic target. In the absence of a solid theoretical background that may build a strategy for the rational design of drugs, the pharmaceutical industry has applied the knowledge of inorganic and organic chemistry for the arbitrary and rather randomized modification of metabolic intermediaries by replacing hydrogen atoms in a model molecule with any other element or compound. This approach has been successful in the battle against many diseases. However, in many other instances such an approach has been unsuccessful.

The era of rational drug design probably started in the 50s when Hans Krebs proposed that, after having an exact description of a metabolic pathway, the "pacemaker" enzyme or "rate-limiting step" had to be identified. This approach certainly decreased the amount of intermediaries to be chemically modified, focusing only on the substrates, products, and allosteric effectors of the "rate-limiting step," instead of dispersing efforts on all the metabolic pathway intermediates. The experimental approaches used in the identification of the pacemaker, key enzymes, "bottlenecks." limiting steps, or regulatory enzymes [1, 2] were

(i) inspection of the metabolic pathway architecture: due to cell economy and for reaching the highest efficiency, pathway control must reside in the enzymes localized at the beginning of a pathway or after a branch (teleological approach);

(ii) determination of nonequilibrium reactions: those reactions in which the quotient between the mass action ratio ($\Gamma$) and its equilibrium constant (Keq) is low, $\Gamma/\text{Keq} \ll 1$ (thermodynamic approach);

(iii) identification of the steps with the lowest maximal rates ($V_{\max}$) in cellular extracts: the key enzyme of the pathway is the one that has the lowest rate (kinetic approach);

(iv) enzymes with sigmoidal kinetics: steps that are susceptible to alteration in their kinetic properties by compounds different from substrates and products and which may coordinate the entire metabolism (NADH/NAD$^+$; NADPH/NADP$^+$, ATP/ADP; acetyl CoA/CoA; Ca$^{2+}$/Mg$^{2+}$; high pH/low pH) or at least two metabolic pathways (citrate, Pi, AMP, malonyl-CoA);

(v) crossover theorem. Comparing the intermediary concentrations between a basal and an active steady-state pathway flux, the rate-limiting step in the basal condition will be that for which its substrate concentration diminishes and its product concentration increases when the system changes from the basal to the active state or vice versa (crossover point on a histogram of each intermediary versus its normalized variation in concentration);

(vi) the shape of the metabolic flux inhibition curve: a sigmoidal curve on a plot of inhibitor concentration versus flux shows that the sensitive step to the inhibitor exerts no control, that is, there is not proportionality between enzyme activity inhibition and pathway flux inhibition because there is an "excess" of enzyme. On the other hand, a hyperbolic curve indicates that the enzyme susceptible to the inhibitor controls the flux.

## 2. CONTROLLING SITES IN A METABOLIC PATHWAY

Once a site in a metabolic pathway has been identified with at least one of the criteria described above as "the rate-limiting step," researchers have frequently concluded that such enzyme or transporter is the only limiting step of the metabolic flux and extend this conclusion to all cell types and to all conditions.

For example, inspection of the glycolytic pathway (teleological approach) suggests that hexokinase (HK) and phosphofructokinase-1 (PFK-1) (which are at the beginning and after a branch of the pathway) are the key steps of glycolysis. However, all studies on glycolysis in the 60s, 70s, and 80s were performed by taking into account only the intracellular reactions from HK to LDH (i.e., without including the glucose transport reaction through the plasma membrane) and by considering glycolysis as a linear pathway without branches. To this regard, it is recalled that the glucose transporter (GLUT) includes a family of proteins and genes that are susceptible of regulation. Thus, if the extracellular glucose is considered as the initial glycolytic substrate, then another potential key step would be GLUT. Hence, if all the branches of the pathway are considered (Figure 1), then according to the teleological approach there will be additional potential rate-limiting sites.

Application of the thermodynamic and kinetic approaches to glycolysis reveals that HK, PFK-1, and pyruvate kinase (PYK) are the rate-limiting steps because in the living cell they catalyze reactions that are far away from equilibrium ($\Gamma/\text{Keq} = 10^{-3}$–$10^{-4}$), and they are also the slowest enzymes in the pathway by at least one order of magnitude (they have the lowest $V_{\max}$ values).

The use of the enzyme cooperativity approach has established that the regulatory steps of glycolysis are (i) PFK-1 and PYK because they are allosteric enzymes and (ii) HK because it is inhibited by its products (G6P and ADP, or AMP as an ADP-analogue). The application of the crossover theorem (approach no. v) to glycolysis has shown a consistent variation in the PFK-1 substrate (F6P) and product (F1,6BP). Up to now, there are few studies on control of glycolysis using the shape of the inhibitor titrating curve (approach no. vi), due to the lack of specific inhibitors for any of the three presumed key steps. An exception is iodoacetate which is indeed a potent inhibitor of GAPDH, but also of other highly reactive cysteine-containing enzymes [3–5]. By using iodoacetate as specific inhibitor, both GAPDH activity and flux showed identical titration curves, leading to the conclusion that GAPDH was the rate-limiting step of glycolysis in *Streptococcus lactis* and *S. cremoris* [6] (see, however, Section 3.2; Glycolysis in lactobacteria below).

All together, these results constitute the main reason why many intermediary metabolism researchers, including the authors of biochemistry text books, have proposed HK, PFK-1, and PYK as the rate-limiting steps of glycolysis. In consequence, to vary the glycolytic flux, one of these enzymes has to be modified.

Although the above-described experimental approaches are qualitative, full control has been automatically assigned to the "key" steps because the concept of the rate-limiting step assumes that there is only one single enzyme controlling the metabolic pathway flux (and the concentration of the final product of the pathway) and, in consequence, assigns values of zero to the control exerted by the other enzymes and transporters. However, as analyzed for glycolysis, researchers have commonly "identified" more than one limiting step. In the case of oxidative phosphorylation (OXPHOS), in the 70s and 80s some researchers considered cytochrome c oxidase as the rate-limiting step, whereas others preferred the ATP/ADP translocator or the Krebs cycle Ca$^{2+}$-sensitive dehydrogenases (for a review, see [7]).

Rephrasing the initial question, which could be the aim of manipulating a metabolic pathway such as glycolysis,

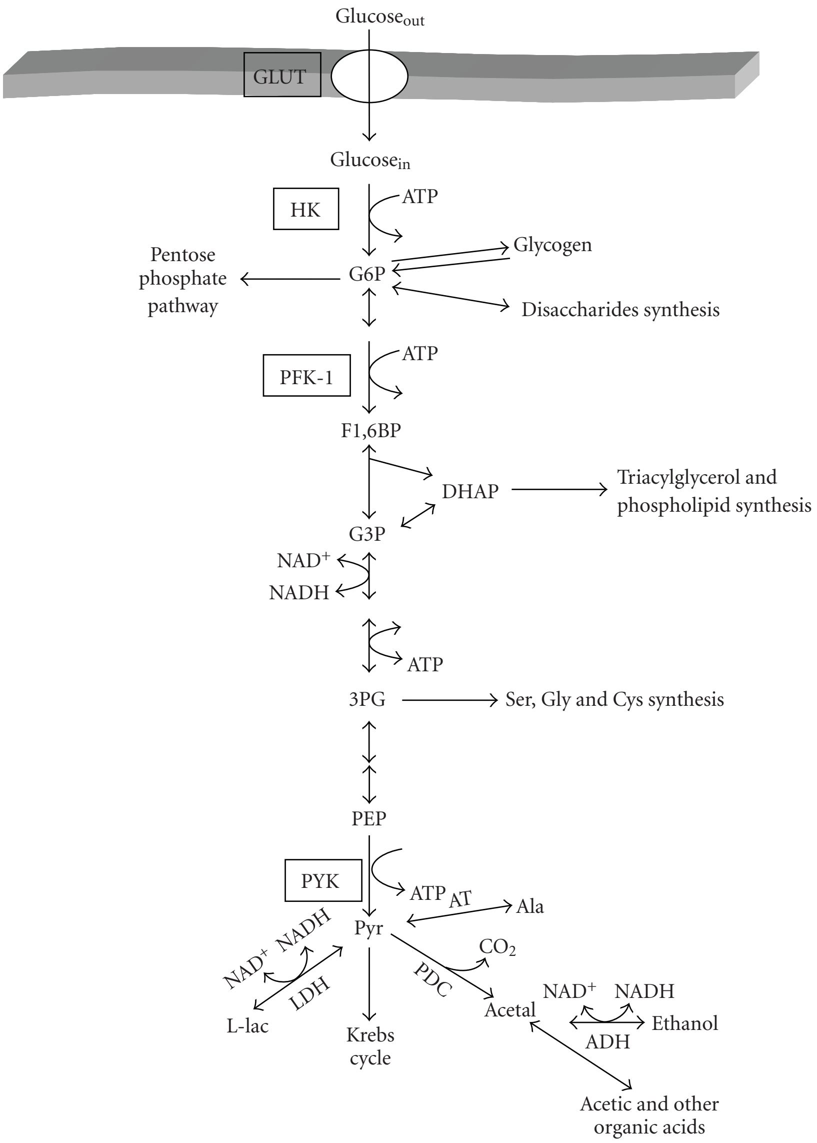


Figure 1: Glycolytic pathway and principal branches. GLUT, glucose transporter; HK, hexokinase; PFK-1, phosphofructokinase-1; G6P, glucose-6-phosphate; F1,6BP, fructose 1,6 bisphosphate; DHAP, dihydroxyacetone phosphate; G3P, glyceraldehyde-3-phosphate; 3PG, 3-phosphoglycerate; PEP, phosphoenolpyruvate; pyr, pyruvate; PYK, pyruvate kinase; L-lac, L-lactate; acetal, acetaldehyde; AT, alanine transaminase. *S. cerevisiae* lacks the LDH gene.

knowing its universal distribution in the living organisms? From a clinical standpoint, the inhibition of glycolysis is relevant for the treatment of human parasitic or pathological diseases such as cancer. The glycolytic reactions are almost identical in all organisms; in addition, the enzymes catalyzing these reactions are highly conserved throughout the evolutionary scale (their amino acid sequences are highly similar). In mammals, the genes of the 12 glycolytic enzymes are scattered throughout the genome, generally in different chromosomes, whereas in bacteria many of the glycolytic enzymes are clustered in operons [8]. However, there are organisms (like some human parasites) that contain enzymes with remarkable differences in their biochemical properties (substrate selectivity, catalytic capacity, stability, and oligomeric structure), or in genetic expression regulation in comparison to the human enzymes, which could be considered as drug targets.

Furthermore, some glycolytic products are of commercial interest such as ethanol for wine, beer, and other alcoholic beverages; $CO_2$ for bread manufacturing; and lactic acid and other organic acids for cheese production. Thus, from a biotechnological standpoint, it is convenient to accelerate the pathway flux to diminish the processing time and it is also desirable to increase the concentration of the metabolite to obtain robust commercial products. Here, it is important to emphasize that the metabolic pathways are designed to attain changes in flux with minimal disturbances in the intermediary concentrations. For example, the glycolytic flux in skeletal muscle can increase from rest to an active state by 100 fold, without large changes in

Table 1: Overexpression of glycolytic enzymes in different cell types.

| Cell type | Enzyme | Activity (overexpression fold) | Flux (% Control) | Reference |
|---|---|---|---|---|
| *Saccharomyces cerevisiae* | HK | 13.9 | 107 | [12] |
| | PFK-1 | 3.5, 3.7,5 | 102 | [9, 10, 12] |
| | PYK | 8.6 | 107 | [12] |
| | PDC | 3.7 | 85 | [13] |
| | ADH | 4.8 | 89 | [12] |
| | PFK-1 + PYK | 5.6 + 1.3 | 107 | [12] |
| | GAPDH + PGK + PGAM + ENO + PYK + PDC + ADH | 1.4 + 1.7 + 16 + 4 + 10.4 + 1.08 + 1.4 | 121 | [12] |
| | GAPDH + PGK + PGAM + ENO + PYK + PDC + ADH | 1.5 + 1.4 + 3.4 + 1.5 + 2.5 + 1.1 + 1.2 | 94 | [11, 14] |
| *Escherichia coli* | PFK | 8.7 | 72 | [15] |
| | PYK | 2.9, 4.2 | 91,95 | [16] |
| *Lactococcus lactis* | GAPDH | 14-210 | 100 | [17] |
| *Aspergillus niger* | PFK | 3 | 100 | [18] |
| | PYK | 5 | 100 | |
| *Chinese hamster ovary* | PFK | 2.2, 3.4, 3.7 | 100 | [19] |

Flux to ethanol was for *S. cerevisiae* and *E. coli*; flux to citrate was for *A. niger*; and flux to L-lactate was for hamster.

metabolites. Then, it is physiologically more common to change a metabolic flux and the production of the final metabolite in the pathway than varying the intermediary concentrations [2]. However, we will see that, by using a suitable approach of metabolic control analysis, it is possible to design strategies to manipulate not only fluxes but also metabolic intermediary concentrations.

## 3. IN VIVO OVEREXPRESSION EXPERIMENTS OF ENZYMES

### *3.1. Glycolysis in yeasts*

When the yeast *Saccharomyces cerevisiae* is exposed to high glucose (>2%; 0.11 M), the genes of all glycolytic enzymes are induced (PDC and ENO increase their expression by 20 fold; PGK, PYK, and ADH, 3–10 times; and the others, 2 fold in average) [8–11]. However, when the methodological development of genetic engineering allowed modulating the expression of enzymes within cells, researchers turned to the rate-limiting step concept to manipulate a metabolic pathway to increase flux and/or its intermediates, hypothesizing that the overexpression of only one, or of a few key glycolytic genes, should increase the flux.

Historically, Heinisch [9] in Germany was the first author to obtain a 3.5 fold overexpression of PFK-1 in *S. cerevisiae*, but surprisingly he observed that the rate of ethanol production was not modified. Subsequent experiments for increasing the ethanol production rate by overexpressing either each of the presumed limiting steps, or in combination with other glycolytic enzymes (Table 1), have been unsuccessful and, even in some cases, a slight decrease in flux has been attained. For instance, the simultaneous overexpression of seven enzymes of the final section of glycolysis induced only a 21% increase in ethanol production after 2 hours of culture (Table 1) [11]. This was accompanied by a 10–20% decrease in PFK-1 expression, which might have attenuated the flux increase.

In yeasts, HK is not product inhibited by G6P or ADP; instead, it is strongly feedback inhibited by trehalose-6-phosphate (Tre6P). This metabolite is synthesized from G1P by Tre6P synthase and Tre6P phosphatase. Deletion of the Tre6P synthase gene does not bring about an increased ethanol production, but it rather induces a defective cellular growth on glucose and fructose and a lowered ethanol production, as a result of a highly active HK that leads to hyperaccumulation of hexose phosphate metabolites (particularly F1,6BP) and fast depletion of ATP, Pi, and downstream metabolites. The explanation for this event is that, in the Tre6P synthase mutants, the rate of glucose phosphorylation exceeds the rate of glycolytic ATP synthesis (named "turbo effect"). Heterologous expression of a Tre6P-insensitive HK does not recover completely the wild-type phenotype. Furthermore, deletion of the Tre6P synthase gene in the Tre6P-insensitive HK strain did affect growth, suggesting other interactions and functions of Tre6P synthase in the control of sugar metabolism, at least in *Schizosaccharomyces pombe* [20].

Davies and Brindle [10] obtained a 5-fold overexpression of PFK-1 in *S. cerevisiae*, but the increase in ethanol production was not attained under anaerobic conditions. There was a slight increase in ethanol production in resting cells in aerobic conditions, under which the mitochondrial metabolism contributes to the ATP supply. In all these works, it may be noted that enzyme overexpression indeed affects the concentration of several intermediaries, but this effect has not been further examined.

It is worth noting that the experiments described in Table 1 do not rigorously reproduce the physiological situation, in which overexpression of all the enzymes should be carried out in the proportions found in the organisms. The rationale behind this observation is that overexpression of only one "limiting" step leads to a flux control redistribution, a condition at which other steps now become rate limiting. Thus, the concept of "rate-limiting step" offers no simple answer to the question of increasing the yeast glycolytic flux, and it rather makes this problem to appear as a difficult task to solve. In contrast, it seems that all relevant controlling steps have to be overexpressed, thus reproducing what natural selection has already successfully accomplished.

In addition to *S. cerevisiae*, overexpression of glycolytic enzymes in other organisms such as *E. coli* [15, 16], lactobacteria [17], tomato [21], potato [22], and hamster ovary cells [19] has been accomplished, although without increasing flux (Table 1). It is somewhat surprising to note that in the glycolytic enzyme overexpression experiments, the strong inhibitory effect of G6P (or Tre6P in *S. cerevisiae*), and citrate on HK and PFK-1, respectively, have been neglected. This regulatory mechanism does not disappear in the cells overexpressing the enzymes but, on the contrary, it is exacerbated. Then, what would be the aim of overexpressing HK, PFK-1 or any other allosteric, or strongly product-inhibited enzyme if they will be more inhibited?

A successful experiment of increasing the glycolytic flux was performed in primary cultures of rat hepatocytes [23]. HK and glucokinase (GK) were overexpressed by using adenovirus as carrier. The transformed hepatocytes showed higher activity of 18.7- and 7.1-times for HK and GK, respectively, at 3 mM glucose, and of 6.3- and 7.1-times at 20 mM glucose. However, at 20 mM glucose, the flux to lactate was not modified in HK-transformed cells, just like the experiments described above (Table 1). In contrast, with GK overexpression, a 3-fold increase in flux was achieved. The mechanistic difference is the HK inhibition by G6P (10 mM G6P inhibits HK activity by 90%), whereas GK is not product inhibited.

### 3.2. Glycolysis in lactobacteria

*Lactococcus lactis* is used in cheese production. For this purpose, *L. lactis* ferments lactose to lactic acid by glycolysis. The end products, lactate and $H^+$, are expelled and acidify the external medium which contributes to cheese flavor and texture and inhibits the growth of other bacteria. Similarly to yeast, the lack of carbon source in lactobacteria promotes a metabolic change that leads to the production of formic and acetic acids, ethanol, and, in a lower proportion, L-lactic acid, altering the product quality. Thus, from a commercial point of view, it does not seem important to know what controls the flux to lactate (because its rate of production is adequate), but what controls the branching flux.

To understand the process, and to eventually inhibit the production of secondary acids, Andersen et al. [24] constructed LDH mutants, using a synthetic promoter library for tuning the gene expression. In mutants lacking this enzyme, most of the pyruvate was transformed into acetic and formic acids (Figure 1). In turn, flux to lactate was affected in mutants expressing only 10% or less of wild-type LDH levels, which indicated that LDH exerts no control of the glycolytic flux in wild-type bacteria. Only with a normal content of this enzyme (100%), flux toward secondary acids was prevented. Therefore, the flux to formic and acetic acids is negatively controlled by LDH, and positively by PYK [17, 25]. As in *S. cerevisiae*, overexpression of PFK-1, PYK, or GAPDH in lactobacteria did not increase the flux to L-lactic acid [17, 25]. Similarly to *E. coli* glycolysis [26], glycolysis in *L. lactis* was controlled by the ATP demand when working below its maximum capacity [27, 28], whereas, under high-rate conditions, the glucose and lactate transporters exerted the main flux control [28]. Furthermore, this kind of observations indicates that the flux control may reside outside the pathway [27–29], and it also supports the proposal by Hofmeyr and Cornish-Bowden [30] that the end-product demand (which is usually overlooked in studies of metabolism because these metabolites are frequently not considered as part of the pathway) might be essential in flux control.

### 3.3. Glutathione and phytochelatin synthesis in plants

Glutathione ($\gamma$-Glu-Cys-Gly; GSH) is the most abundant nonproteinaceous thiol compound (1–10 mM) in almost all living cells. GSH is involved in the oxidative stress processing, xenobiotic detoxification, and, in some plants and yeasts, in the inactivation of toxic heavy metals (for a recent revision see [31]). GSH is synthesized by two enzymes: $\gamma$-glutamylcysteine synthetase ($\gamma$-ECS) and glutathione synthetase (GS) (Figure 2), which catalyze reactions with high-equilibrium constants (Keq > 1000). Under a low GSH demand (unstressed conditions), the producing block of enzymes has to receive information from the last part of the pathway to (i) avoid the excessive and toxic accumulation of the intermediary $\gamma$-EC and (ii) reach a stable steady state [32]. This information transfer is mediated by GSH, which exerts strong competitive inhibition of $\gamma$-ECS [33] (Figure 2). GSH and Cys also exert inhibition on the ATP-sulfurylase (ATPS) and on sulfate transporters (Figure 2) (for a review, see [31]). The feedback inhibition of $\gamma$-ECS has led several researchers to propose that this enzyme is the rate-limiting step of GSH synthesis [33–35]. Although there are no studies about the pathway's behavior under stressed conditions, which means under a high GSH demand, the proposal that $\gamma$-ECS is the key enzyme has been automatically extended to any environmental condition such as heavy metal exposure.

By assuming that $\gamma$-ECS is the rate-limiting step, many research groups have tried to increase, in plants and yeasts, the rate of synthesis and the concentration of GSH and phytochelatins (PCs) with the aim of fortifying their heavy metal resistance and storage capacity, mainly toward $Cd^{2+}$. The development of organisms able to grow in soils and water systems contaminated with heavy metals, which may

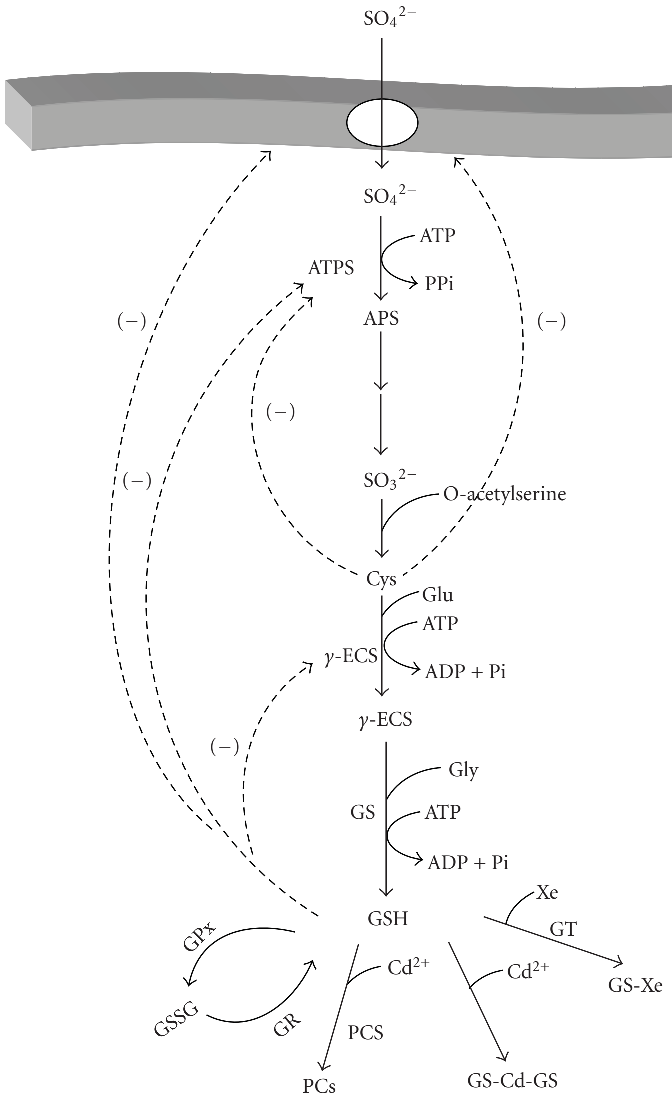


Figure 2: Sulfur assimilation and glutathione and phytochelatins synthesis in plants ATPS, ATP sulfurylase; APS, adenosine 5′ phosphosulphate; γ-ECS, γ-glutamyl cysteine synthetase; γ-EC, γ-glutamyl cysteine; GS, glutathione synthetase; GSH, reduced glutathione; GSSG, oxidized glutathione; GPx, GSH peroxidase; GR, GSH reductase; PCS, phytochelatin synthase; PCs, phytochelatins; GT, GSH-S-transferases; Xe, xenobiotic; GS-Xe, glutathione-xenobiotic complex. The reactions are not shown stochiometrically. GR uses the cofactor NADPH. The $Cd^{2+}$-GSH complex formation (cadmium bis-glutathionate) is fast and spontaneous and does not require enzyme catalysis. Modified from [31].

have the ability of accumulating toxic metal ions, is of biotechnological interest for bioremediation strategies.

With this goal in mind, researchers have then overexpressed γ-ECS and other pathway enzymes, including phytochelatin synthase (PCS) (Table 2). Some of these experiments have been partially successful in increasing GSH levels, although this has been rather marginal with no correlation between enzyme levels and GSH concentration. Unfortunately, these overexpression experiments have not been accompanied by determinations of fluxes or other relevant metabolite concentrations such as PCs or Cys. On the other hand, the overexpression of PCS has surprisingly induced oxidative stress and necrosis instead of increasing $Cd^{2+}$ accumulation and resistance [36]. This result suggests that, under high GSH demand (i.e., for PCs synthesis and for direct heavy metal sequestration by GSH), the GSH concentration does not suffice for maintaining the other essential GSH functions such as oxidative stress management and xenobiotic detoxification.

Another problem in the study of GSH biosynthesis for its eventual manipulation is that the pathway has been analyzed considering only the GSH-synthetic reactions without taking into account the GSH-consuming reactions (Figure 2), [31]. The analysis of an incomplete pathway leads to misleading conclusions about the control of flux. Metabolic modeling has shown that only with the incorporation of the consuming reactions of the pathway end products, a true steady state can be established [30]. In conclusion, without a solid theoretical

Table 2: GSH and phytochelatin synthesis enzymes overexpression in plants and yeasts.

| Overexpressed enzyme (activity fold) | Organism (experimental condition) | Metabolite (increment fold) | Reference |
|---|---|---|---|
| ATP sulfurylase (2.1) | *Brassica juncea* | 2.1 [GSH] | [37] |
| ATP sulfurylase (4.8) | Tobacco (unstressed) | 1.3 [$SO_4^{2-}$] | [38] |
| O-acetyl-serine thiol-lyase (2.5) | Tobacco (unstressed) | 2 [Cys]<br>0 [GSH] | [39] |
| Serine acetyl transferase (>10) | Potato chloroplasts (unstressed) | 2 [Cys]<br>0 [GSH] | [40] |
| *E. coli* GS (90) | *Populus tremula* (unstressed) | 0 [GSH] | [34] |
| GS (3) | *S. cerevisiae* (unstressed) | 0 [GSH] | [41] |
| *E. coli* $\gamma$-ECS (>2) | *Brassica juncea* (unstressed)<br>*B. juncea* (+100 $\mu$M $Cd^{2+}$) | 0 [GSH]<br>4 [GSH][a] | [35] |
| $\gamma$-ECS (2.1) | *S. cerevisiae* (unstressed)) | 1.3 [GSH] | [42] |
| *E. coli* $\gamma$-ECS (50) | *Populus tremula* (unstressed) | 4.6 [GSH] | [34] |
| *E. coli* $\gamma$-ECS (4.9) | *Brassica juncea* (unstressed) *B. juncea* (+200 $\mu$M $Cd^{2+}$) | 3.5 [GSH][b]<br>1.5 [GSH][b] | [43] |
| *E. coli* $\gamma$-ECS (40) | Tobacco (unstressed) | >4 [GSH] | [44] |
| $\gamma$-ECS (9.1) + GS (18) | *S. cerevisiae* (unstressed) | 1.8 [GSH] | [45] |
| PCS (>2) | *Arabidopsis thaliana* (+85 $\mu$M $Cd^{2+}$) | 0 [GSH] | [36] |
| Vacuolar transporter of PC-Cd complexes (>2) | *S. pombe* | Higher $Cd^{2+}$ resistance | [46] |

[a] The increase was only in roots with no effect on shoots. [b] The increase was only in shoots with no effect on roots.

framework, the overexpression of only one enzyme (the "rate-limiting step"), or of many arbitrarily selected enzymes (Tables 1 and 2), the problem of increasing the flux or metabolite concentrations cannot be solved.

### 3.4. Overexpression of proteins from other metabolic pathways

There are some successful examples of the genetic engineering approach to manipulate metabolism:

(i) overexpression (approx. 23 fold) of the five genes of the tryptophan synthesis pathway in *S. cerevisiae*, to increase (9-fold) flux [47];

(ii) increase in amino acids (Trp, Ile, Lys, Val, Thr) and trehalose production in *Corynebacterium glutamicum*, in which some proteins of each metabolic pathway are simultaneously overexpressed, but some of them with mutations that confer insensitivity to feedback inhibition [48–53]. In these transformed bacteria, the end products are indeed overproduced and their excretion is accelerated;

(iii) overexpression of PFK and PyK to increase ethanol production by 35% in *E. coli*, although lactic acid formation was not modified [16];

(iv) mannitol 1-phosphate dehydrogenase and mannitol 1-phosphatase overexpression to increase mannitol production by 27–50% in LDH-deficient *Lactococcus lactis* [54];

(v) increase in sorbitol production (5 fold) in LDH-deficient *Lactobacillus plantarum* through the overexpression of sorbitol 6-phosphate dehydrogenase (activity up to 250 fold in mutants *versus* wild type) [55];

(vi) overexpression of PFK (14 fold) or LDH (3.5 times) to increase 2-3 times the homolactic fermentation flux in *Lactococcus lactis* growing on maltose, and in parallel decrease fluxes toward secondary acids and ethanol [56].

## 4. DOWNREGULATION OF ENZYMES TO MANIPULATE METABOLISM

### 4.1. Glycolysis in tumor cells

Glycolysis is enhanced in human and animal cancer cells (reviewed in [57]). Several glycolytic enzymes are overexpressed in at least 70% of human cancers [58]. Except for glucose transporter 1 (GLUT-1), the other 11 glycolytic enzymes (HK to LDH) are overexpressed in brain and nervous system cancers. Prostate and lymphatic nodule cancers (Hodgkin and non-Hodgkin lymphomas; myelomas) overexpress 10 glycolytic enzymes (except for HK; in prostate cancer GLUT1

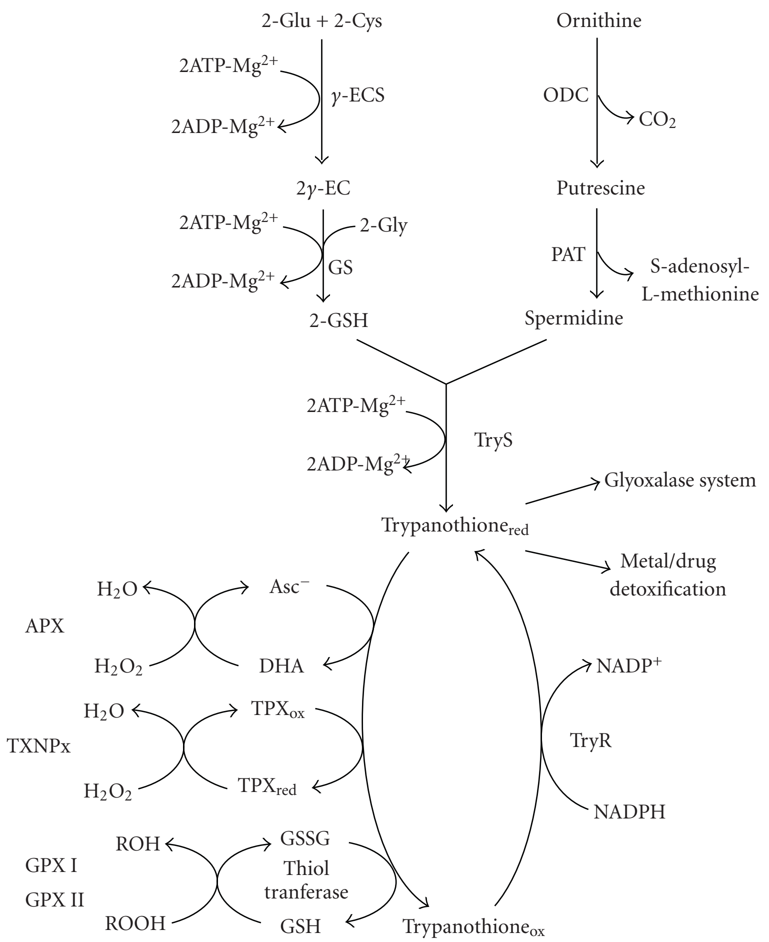


FIGURE 3: Trypanothione synthesis in trypanosomatids. The trypanothione producing enzymes are $\gamma$-ECS, GS, ODC, aminopropyl transferase (PAT), and TryS. The trypanothione consuming enzymes are ascorbate peroxidase (APX); tryparedoxin peroxidases (TXNPx); trypanothione-glutathione thiol transferase (thiol transferase); and glutathione peroxidases I (GPX I) and II (GPX II). The regenerating enzyme is TryR. APX, thiol transferase, and GPX II have only been described in *T. cruzi*. This last parasite lacks ODC activity, but it has developed high-affinity transporters for putrescine, cadaverine, and spermidine [71].

is also overexpressed). There is a second group of cancers that overexpresses 6–8 glycolytic genes (skin, kidney, stomach, testicles, lung, liver, placenta, pancreas, uterus, ovary, eye, head and neck, and mammary gland). A third group includes those cancers overexpressing 1 or 2 glycolytic genes (bone, bone marrow, cervix, and cartilage) [58].

In animals, gene expression of glycolytic enzymes is regulated (both coordinately and individually) under hypoxic conditions by hypoxia-responsive transcription factors such as HIF-1$\alpha$ (hypoxia-inducible factor 1$\alpha$), SP family factors, AP-1, and possibly MRE (metal response elements) [8, 59–61]. HIF-1$\alpha$ is probably the principal coordinator in gene induction. There are binding sites (consensus sequence ACGT) for HIF-1$\alpha$ in the promoters of genes for HK [62], PFK-1, ALDO, GAPDH, PGK, ENO, PYK, and LDH (reviewed in [8]). TPI and perhaps HPI and PGAM are also induced by hypoxia, but it is not clear whether HIF-1$\alpha$ mediates this induction [8], and whether this factor regulates other metabolic pathways associated with glucose catabolism. For example, although glycogen phosphorylase is overexpressed under hypoxia in human tissues [63], the role of HIF-1 has not been demonstrated.

If direct manipulation of pathway genes becomes difficult, then the overexpression or repression of transcription factors such as HIF-1$\alpha$, AP1, and MREs might solve the problem of changing flux, although overexpression of transcription factors may also be difficult due to the numerous upstream and downstream factors involved.

### 4.2. Glycolysis in Trypanosoma brucei

The kinetoplastid parasites *Trypanosoma cruzi, Trypanosoma brucei, and Leishmania* are the causative agents of Chagas disease, African trypanosomiasis, and leishmaniasis, respectively. The available drugs to treat these diseases are highly toxic for humans. Moreover, the parasites may become resistant, and hence the search for new drugs and drug targets is relevant for solving these public health problems.

In these parasites, the metabolism is organized in a peculiar way; they have a subcellular structure called glycosome in which several metabolic pathways take place: gluconeogenesis, reactions of the pentose phosphate pathway, purine salvage and pyrimidine biosynthesis, $\beta$-oxidation of fatty acids, fatty acid elongation, biosynthesis of ether lipids, and the first seven steps of glycolysis. In fact, approximately 90% of glycosome enzyme content corresponds to glycolytic enzymes [64]. Glycosomal glycolytic enzymes have unique structural, kinetic, and regulatory features not found in their human counterparts, and therefore have been the subject of extensive biochemical studies to use them as drug targets [65]. The rationale behind this is to synthesize inhibitors that affect mainly the parasitic enzymes with relatively low effect on the human enzymes since the infective parasite stages rely mostly on glycolysis for ATP supply.

There are reports on the design of presumed specific inhibitors for some of the *T. brucei* glycolytic enzymes: GLUT (bromoacetyl-2-glucose) [66], HK, HPI, PFK, ALDO, TPI, GAPDH, PGK, PYK, and glycerol-3-phosphate dehydrogenase [67]. Although the purified enzymes display very low *Ki* values for these inhibitors and some of them inhibit parasite growth or infective capabilities, their effect on inhibiting the glycolytic flux has not been explored. Therefore, it is not yet possible to directly ascribe the effects seen in parasite culture with the in vitro effects on the isolated enzymes. To identify the best drug targets, determination of the flux control steps of glycolysis in *T. brucei* has been recently initiated [68].

### 4.3. Trypanothione synthesis in kinetoplastid parasites

Trypanothione ($TSH_2$) is a reducing agent present in trypanosomatids that is synthesized from one spermidine and two GSH molecules by $TSH_2$ synthetase (TryS) (Figure 3). This metabolite and its reducing enzyme, $TSH_2$ reductase (TryR), replace the antioxidant and metabolic functions of the more common GSH/GSH reductase system present in mammals. In fact, most of the antioxidant metabolism of these parasites depend on $TSH_2$ (Figure 3) [69, 70]. Thus, the enzymes of this metabolic pathway have been proposed as drug targets for killing the parasites.

Several studies have focused in assessing TryR as drug target. Diminution in its gene transcription yields a loss of activity between 56–90%, depending on the genetic technique [72–75]. In knockdown *T. brucei* cells (i.e., when TryR activity has diminished to less than 10% of the wild-type level), the parasites show growth diminution and higher sensitivity to $H_2O_2$ in culture and loss of infectiveness in mice. However, $TSH_2$ and thiol compound contents were not affected [75]. TryR downregulation by >85% in *Leishmania* species causes inability to survive under oxidative stress inside macrophages [72–74]. In contrast, when TryR is 14- and 10 fold overexpressed in *Leishmania* and *T. cruzi*, respectively, there are no significant differences in $H_2O_2$ susceptibility between control and transfected cells; both types of cells are also equally resistant to the oxidative stress-inducers gentian violet, and nitrofurans [76]. Intriguingly, the cellular levels of $TSH_2$, GSH, and glutathionyl-spermidine, determined in both types of experiments (TryR suppression and overexpression) were similar in control and transformed cells.

Other studies have proposed TryS as an alternative drug target. Knockdown of TryS by siRNA in procyclic *T. brucei* causes (i) viability impairment and arrest of proliferation when $TSH_2$ levels decrease to 15% of the wild-type level, (ii) increased sensitivity to $H_2O_2$ and alkyl hydroperoxides, (iii) damage to the plasma membrane, and (iv) diminution of the $TSH_2$ content and accumulation of GSH and glutathionyl-spermidine [77]. A similar metabolite variation (lower $TSH_2$; higher GSH) was attained with a TryS knockdown induced by siRNA in the bloodstream form of *T. brucei* [78]. This TryS knockdown also induced an increased sensitivity to different compounds that affect $TSH_2$ metabolism such as arsenicals, melarsen oxide, trivalent antimonials, and nifurtimox [78]. Indeed, western blot analysis showed, in addition to the expected (10-fold) decrease in TryS protein, a 2-3-folds increase in $\gamma$-ECS and TryR. The changes in expression of other enzymes suggest unveiled compensatory or pleiotropic effects on $TSH_2$ metabolism.

Other researchers have selected $\gamma$-glutamylcysteine synthetase ($\gamma$-ECS), the presumed rate-limiting step of GSH synthesis, as an alternative drug target of $TSH_2$ synthesis in *T. brucei* (Figure 3). Knockdown of $\gamma$-ECS gene in the parasite induces cell death and depletion of GSH and $TSH_2$ only after 80% decrease in the enzyme content [79]. The $\gamma$-ECS knockdown cells are rescued from death by adding external GSH, which elevates the cellular GSH and $TSH_2$ levels [79].

Glutathione synthetase (GS) has not been manipulated in trypanosomatids, or in any other organism, perhaps because it has been considered as a nonrate-limiting step of GSH and $TSH_2$ biosynthesis. However, DNA microarray analysis of antimonite-resistant *Leishmania tarentolae* shows increased transcription of $\gamma$-ECS, GS, and P-glycoprotein A RNAs [80]. Although it was not evaluated whether increase in gene transcription correlated with an increase in enzyme activity, it may be possible that under high GSH demand (i.e., under oxidative stress conditions) GS might exert control of $TSH_2$ synthesis. On the other hand, ornithine decarboxylase (ODC) overexpression in *T. brucei* (the presumed limiting step of spermidine synthesis) causes no change in $TSH_2$ levels [81]. Therefore, ODC does not seem to be a controlling step of $TSH_2$ synthesis.

Although almost full inhibition (>80%) of gene transcription or activity of any of these enzymes results in parasite death, the question remains of how $TSH_2$ metabolism is affected when the enzymes are less inhibited. For example, in the therapeutic treatment of patients it is certain that drugs have to be administered for long periods of time. If the parasites are not completely cleared from the patient, disease recurrence and generation of drug-resistant parasites are possible. The results described above indicate that each enzyme by itself has low control on $TSH_2$ synthesis and concentration; therefore, highly specific and very potent inhibitors have to be designed in order to attain the required full activity blockade to affect $TSH_2$ metabolism in these parasites.

## 5. THEORY OF METABOLIC CONTROL ANALYSIS

The metabolic control analysis (MCA) was initially developed by Kacser and Burns in Scotland [82, 83] and by Heinrich and Rapoport in East Germany [84, 85]. This analysis establishes a theoretical framework that explains the results observed with the enzyme overexpression and downregulation experiments. In addition, it helps to identify and design experimental strategies for the manipulation of a given process in an organism (heavy metal hyperaccumulation; increased production of ethanol, $CO_2$, lactate or acetate; or inhibition of a metabolic pathway flux with therapeutic purposes). MCA rationalizes the quantitative determination of the degree of control that a given enzyme exerts on flux and on the concentration of metabolites. Different experimental approaches have been developed to detect and direct what has to be done and measured, in order to identify and understand why an enzyme exerts a significant or a negligible control on flux and metabolite concentration in a metabolic pathway. Thus, the application of this analysis avoids the "trial and error" experiments for identifying and manipulating the conceptually wrong "rate-limiting step."

To understand how a metabolic pathway is controlled and could be manipulated, its control structure has to be evaluated. The control structure of a pathway is constituted by the flux control coefficient ($C^J_{vi}$), which is the degree of control that the rate ($v$) of a given enzyme $i$ exerts on flux $J$; the concentration control coefficient ($C^X_{vi}$), which is the degree of control that a given enzyme $i$ exerts on the concentration of a metabolite ($X$); and the elasticity coefficients. The control coefficients are systemic properties of the pathway that are mechanistically determined by the elasticity coefficients ($\varepsilon^{vi}_X$), which are defined as the degree of sensitivity of a given enzyme $v_i$ (i.e., the enzyme's ability to change its rate) when any of its ligands ($X$: substrate, products or allosteric modulators) is varied.

The flux control coefficient is defined as

$$C^J_{v_i} = \frac{dJ}{dv_i} \bullet \frac{v_{io}}{J_o}, \qquad (1)$$

in which the expression $dJ/dv_i$ describes the variation in flux ($J$) when an infinitesimal change is done in the enzyme $i$ concentration or activity. In practice, the infinitesimal changes in $v_i$ are undetectable, and hence measurable noninfinitesimal changes are undertaken. If a small change in $v_i$ promotes a significant variation in $J$, then this enzyme exerts an elevated flux control (Figure 4, position 1). In contrast, if a rather small or negligible change in flux is observed when $v_i$ is greatly varied, then the enzyme does not exert significant flux control (Figure 4, position 2). To obtain dimensionless and normalized values of $C^J_{vi}$ the scaling factor $v_{io}/J_o$ is applied, which represents the ratio between the initial values from which the slope $dJ/dv_i$ is calculated. If all $C^J_{vi}$ of the pathway enzymes and transporters are added up, the sum comes to one (summation theorem).

The MCA clearly distinguishes between the control exerted by a given enzyme on flux (flux control coefficient) and on the metabolite concentration (concentration control coefficient). Thus, an enzyme can have significant control on a metabolite concentration but not on the pathway flux. This distinction is important for biotechnology purposes. On one hand, the use of the rate-limiting step concept for manipulating metabolic pathways does not make such differentiation, which probably has contributed to the many unsuccessful experiments reported in the literature; on the other hand, it should be clearly defined whether the aim of the project is to increase flux and/or a metabolite concentration since MCA establishes for each aim a different experimental design.

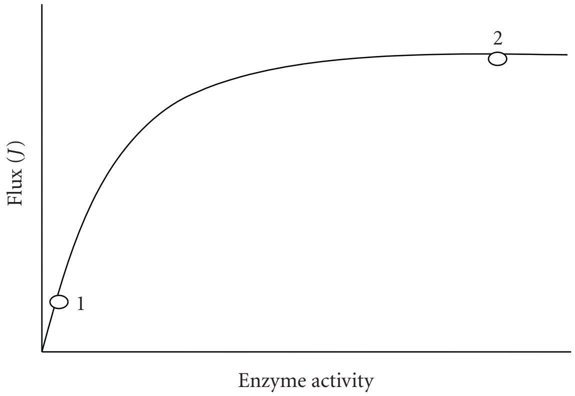


FIGURE 4: Experimental determination of flux control coefficient.

To determine the flux control coefficient of a given enzyme, small variations in the enzyme content, or preferentially, in activity are required, without altering the rest of the pathway, and then the changes in flux are determined. The experimental points are plotted as shown in Figure 4 to calculate the slope at the reference point $v_{io}/J_o$. This experiment, apparently easy to perform, has demanded great intellectual and experimental effort. Several experimental strategies have been developed to determine $C^J_{vi}$:

(i) formation of heterokarionts and heterocygots (classical genetics),

(ii) titration of flux with specific inhibitors,

(iii) elasticity analysis,

(iv) mathematical modeling (in silico biology),

(v) in vitro reconstitution of metabolic pathways,

(vi) genetic engineering to manipulate in vivo protein levels.

### 5.1. Classical mendelian genetics

The arginine biosynthesis in *Neurospora crassa* was the first metabolic pathway in which flux control coefficients were experimentally determined by Kacser's laboratory [86]. This fungus forms multinucleated mycelia that facilitate the generation of polyploid cells. By mixing different ratios of spores containing genes encoding wild (active) and mutant (inactive) enzymes of this pathway, it was possible to generate heterokaryont mycelia with different content, and activity, of four pathway enzymes. The authors built plots of enzyme activity versus flux (see Figure 4) for acetyl-ornithine aminotransferase, ornithine transcarbamoylase, arginine-succinate synthetase, and arginine-succinate lyase. All the experimental points of these heterokaryonts localized near to position 2 of Figure 4 with $C_{vi}^{J\,\mathrm{arg}} = 0.02$–$0.2$ (flux control by these enzymes was only 2–20%), which indicated that none of these enzymes exerted significant control on arginine synthesis. The authors did not determine the remaining flux control (75%), which might reside in carbamoyl-phosphate synthetase I (this mitochondrial ammonium-dependent isoform can be bound to the mitochondrial inner membrane or form complexes with ornithine transcarbamoylase [87, 88]) and in mitochondrial citruline/ornithine transporter, both of which have been proposed as limiting steps, or might be in the arginine demand for protein synthesis.

Organisms with many alleles of one enzyme may form homo-and heterozygotes expressing different activity levels. *Drosophila melanogaster* has three ADH alleles encoding for isoforms with different $V_{\max}$. When three natural homozygotes, a null mutant, and some heterozygotes were generated, different ADH activities were attained but the ethanol consuming rate did not change (Figure 4, position 2). It was concluded that the ADH flux control was near zero [89].

### 5.2. Titration of flux with inhibitors (control of oxidative phosphorylation)

Oxidative phosphorylation (OXPHOS) is the only pathway for which specific and potent inhibitors for many enzymes and transporters are available. OXPHOS is divided in two segments (Figure 5): the oxidative system (OS) formed by substrate transporters (pyruvate, 2-oxoglutarate, glutamate, glutamate/aspartate, dicarboxylates), Krebs cycle enzymes, and the respiratory chain complexes; and the phosphorylating system (PS) constituted by the ATP/ADP (ANT) and Pi (PiT) transporters, and ATP synthase. The proton electrochemical gradient $(\Delta\mu^{-}{}_{\mathrm{H}}{}^{+})$ connects the two systems.

When the flux (ATP synthesis) is titrated by adding increasing concentrations of each specific inhibitor, plots are generated in which the enzyme activity is progressively diminished by increasing inhibitor concentration. Hence, the $C_{vi}^{J}$ value depends on the type of inhibitor used

(a) for irreversible inhibition,

$$C_{v_i}^{J} = \left(\frac{-I_{\max}}{J_o}\right)\left(\frac{dJ}{dI}\right)_{[I]\to 0}, \tag{2}$$

(b) for simple noncompetitive inhibition,

$$C_{v_i}^{J} = \left(\frac{-Ki}{J_o}\right)\left(\frac{dJ}{dI}\right)_{[I]\to 0}, \tag{3}$$

(c) for simple competitive inhibition,

$$C_{v_i}^{J} = \left(\frac{-Ki[(1+S)/Km]}{J_o}\right)\left(\frac{dJ}{dI}\right)_{[I]\to 0}, \tag{4}$$

where $J_o$ is the pathway flux in the absence of inhibitor; $I_{\max}$, minimal inhibitor concentration to reach maximal flux inhibition; $Ki$, inhibition constant; $S$, substrate concentration; $Km$, Michaelis-Menten constant; and $dJ/dI$, initial slope $([I] = 0)$ of inhibition titration curve.

To estimate flux control coefficients from inhibitor titration of ADP-stimulated (state 3) respiratory rates (i.e., mitochondrial $O_2$ consumption coupled to ATP synthesis), (2) for irreversible inhibitors was used because researchers assumed that mitochondrial inhibitors such as rotenone, antimycin, carboxyatractyloside, and oligomycin were "pseudoirreversible," due to the enzyme's high affinity for them. However, under this assumption flux control coefficients were usually overestimated [90, 91]. To solve this problem, Gellerich et al. [92] developed (5) for noncompetitive tightly-bound inhibitors and, by using nonlinear regression analysis, it was possible to include all experimental points from the titration curve thus increasing accuracy in calculating $C_{vi}^{J}$:

$$\begin{gathered} J = \left[\frac{n(J_o - J_i)^2 \bullet E^n}{C_o \bullet J_o \bullet E_o{}^n[(n - C_o)\bullet J_o - (n \bullet J_i)\bullet E^n]}\right] + J_i \\ E^2 + (Kd + I - E_o)\bullet E - Kd \bullet E_o = 0, \end{gathered} \tag{5}$$

in which $J_o$ and $J_i$ are the respiration fluxes in the noninhibited $(E = E_o)$ and inhibited $(E = 0)$ states; $Kd$ is the dissociation constant of the enzyme-inhibitor complex, and $n$ is an empirical component that expresses the relationship between substrate concentration and the reaction catalyzed by the enzyme $E$.

The analysis of data in Table 3 shows that OXPHOS is not controlled by only one limiting step, but the flux control is rather distributed among several enzymes and transporters. It is worth noting that the value of the flux control coefficient depends on the content of enzyme or transporter, which varies from tissue to tissue. Perhaps the ATP/ADP translocase in AS-30D hepatoma mitochondria might reach the status of being the "OXPHOS limiting step" with a $C_{\mathrm{ANT}}^{\mathrm{JOxPhos}} = 0.70$, or the Pi transporter in kidney mitochondria [93], or the ATP/ADP translocase and the respiratory chain complex 3 in liver mitochondria [94], but it should be noted that other steps also exert significant control (Table 3). Although the distribution of control varies between tissues, the flux control mainly resides in the PS of organs with high ATP demand such as the heart ($C_{\mathrm{PT+ANT+ATPsynthase}}^{\mathrm{JOxPhos}} = C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.73$), kidney ($C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.75$; $C_{\mathrm{OS}}^{\mathrm{JOxPhos}} = 0.31$), and fast-growing tumors ($C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.98$). In contrast, in the liver ($C_{\mathrm{OS}}^{\mathrm{JOxPhos}} = 0.80$; $C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.65$) and brain ($C_{\mathrm{OS}}^{\mathrm{JOxPhos}} = 0.35$; $C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.41$), the control is shared by both systems.

The situation in skeletal muscle appears controversial. Wisniewski et al. [97] determined that the OXPHOS control was shared by the PS ($C_{\mathrm{PS}}^{\mathrm{JOxPhos}} = 0.62$) and the ATP demand (purified ATPase). In turn, Rossignol et al. [95] concluded that the OS exerted the main control ($C_{\mathrm{OS}}^{\mathrm{JOxPhos}} = 0.68$), but these authors apparently used low-quality mitochondria (low respiratory control values that lead to low rates of ATP synthesis associated with high rates of respiration) that were

Table 3: Control distribution of oxidative phosphorylation.

| Enzyme | $C_{vi}^{JATP}$ | Rat organ mitochondria | Specific inhibitor | Inhibition mechanism | Reference |
|---|---|---|---|---|---|
| NADH-CoQ-oxidoreductase (Site 1 of energy conservation or Complex I of respiratory chain) | 0.15 | Heart (0.5 mM pyr + 0.2 $\mu$M $Ca^{2+}$) | Rotenone | Noncompetitive tightly bound | [93] |
| | 0.26 | Heart (10 mM pyr + 10 mM mal) | | | [95] |
| | 0.31 | Kidney (0.5 mM pyr + 0.2 $\mu$M $Ca^{2+}$) | | | [93] |
| | 0.06 | Kidney (10 mM pyr + 10 mM mal) | | | [95] |
| | 0.06–0.10 | Brain (0.05 mM pyr + 0.4 $\mu$M $Ca^{2+}$) | | | [91] |
| | 0.25 | Brain (10 mM pyr + 10 mM mal) | | | [95] |
| | 0 | Tumor (10 mM glut + 3 mM mal) | | | [96] |
| | 0.27 | Liver (10 mM pyr + 10 mM mal) | | | [95] |
| | 0.13 | Skeletal muscle (10 mM pyr + 10 mM mal) | | | [95] |
| CoQ.cytochrome c oxidoreductase (Site 2 of energy conservation or Complex III of respiratory chain) | 0.01 | Heart | Antimycin | Noncompetitive tightly bound | [93] |
| | 0.19 | Heart | | | [95] |
| | 0.02 | Kidney | | | [95] |
| | 0.05–0.11 | Brain | | | [91] |
| | 0.02 | Brain | | | [95] |
| | 0 | Tumor | | | [96] |
| | 0.43 | Liver (5 mM Succ + 1 $\mu$M $Ca^{2+}$) | | | [94] |
| | 0.07 | Liver | | | [95] |
| | 0.22 | Skeletal muscle | | | [95] |
| Cytochrome c oxidase (Site 3 of energy conservation or Complex IV of respiratory chain) | 0.11 | Heart | Cyanide or azide | Noncompetitive simple | [93] |
| | 0.13 | Heart | | | [95] |
| | 0.04 | Kidney | | | [95] |
| | 0.02–0.07 | Brain | | | [91] |
| | 0.02 | Brain | | | [95] |
| | 0.04 | Tumor | | | [96] |
| | 0.23 | Liver | | | [94] |
| | 0.03 | Liver | | | [95] |
| | 0.20 | Skeletal muscle | | | [95] |
| ATP/ADP transporter (adenine-nucleotides or ATP/ADP transporter, carrier or exchanger) | 0.24 | Heart | Carboxy-atractyloside (CAT) | Noncompetitive tightly bound | [93] |
| | 0.04 | Heart | | | [95] |
| | 0 | Kidney | | | [93] |
| | 0.07 | Kidney | | | [95] |
| | 0.08 | Brain | | | [91] |
| | 0.08 | Brain | | | [95] |
| | 0.60–0.70 | Tumor | | | [96] |
| | 0.48 | Liver | | | [93] |
| | 0.01 | Liver | | | [93] |
| | 0.37 | Skeletal muscle (10 mM Glut + 3 mM mal) | | | [97] |
| | 0.08 | Skeletal muscle | | | [95] |

Table 3: Continued.

| Enzyme | $C_{vi}^{JATP}$ | Rat organ mitochondria | Specific inhibitor | Inhibition mechanism | Reference |
|---|---|---|---|---|---|
| ATP synthase | 0.34 | Heart | Oligomycin | Noncompetitive tightly bound | [93] |
| | 0.12 | Heart | | | [95] |
| | 0.32 | Kidney | | | [93] |
| | 0.27 | Kidney | | | [95] |
| | 0.09–0.20 | Brain | | | [91] |
| | 0.26 | Brain | | | [95] |
| | 0.28 | Tumor | | | [96] |
| | 0.05 | Liver | | | [94] |
| | 0.20 | Liver | | | [95] |
| | 0.10 | Skeletal muscle | | | [97] |
| | 0.10 | Skeletal muscle | | | [95] |
| Pi transporter | 0.15 | Heart | Mersalyl | Noncompetitive simple | [93] |
| | 0.14 | Heart | | | [95] |
| | 0.43 | Kidney | | | [93] |
| | 0.28 | Kidney | | | [95] |
| | 0.13 | Brain | | | [91] |
| | 0.26 | Brain | | | [95] |
| | 0 | Tumor | | | [96] |
| | 0.05–0.12 | Liver | | | [94] |
| | 0.26 | Liver | | | [95] |
| | 0.15 | Skeletal muscle | | | [97] |
| | 0.08 | Skeletal muscle | | | [95] |
| Pyruvate transporter | 0.15 | Heart | $\alpha$-cyano-4-hydroxy-cinnamate | Noncompetitive simple | [95] |
| | 0.03 | Kidney | | | [95] |
| | 0.08 | Brain | | | [91] |
| | 0.26 | Brain | | | [95] |
| | 0.21 | Liver | | | [95] |
| | 0.20 | Skeletal muscle | | | [95] |
| Dicarboxylates transporter | 0.05–0.14 | Liver | Malate or butyl-malonate | Competitive simple | [94] |
| External ATPase | 0.40 | Skeletal muscle | Purified ATPase addition | | [94] |

not incubated under near physiological conditions (10 mM pyruvate, 10 mM malate, 10 mM Pi, pH 7.4 in Tris buffer), and the authors incorrectly assumed that rotenone and antimycin were irreversible inhibitors. It is notorious that in all works shown in Table 3 at least one of these mistakes is evident.

There are some inhibitors for enzymes and transporters from other pathways, but they are not quite specific and may affect other sites. Due to the fact that there are no inhibitors for every step in these pathways, only one flux control coefficient has been determined by inhibitor titration. Examples of these inhibitors are 6-chloro-6-deoxyglucose for glucose transporters in bacteria, 2-deoxyglucose for HPI, iodoacetate for GAPDH [6], 1,4-dideoxy-1,4-imino-D-arabinitol for glycogen phosphorylase [98], oxalate and oxamate for LDH, 6-amino nicotinamide for the phosphate pentose pathway [99], amino-oxyacetate for aminotransferases and kirureninase (tryptophan synthesis), norvaline for ornithine transcarbamylase, mercaptopycolinate for PEP carboxykinase, acetazolamide for carbonic anhydrase, and isobutyramide for ADH (compiled by Fell [2]).

### *Potential uses of the experimental approach*

Mitochondrial pathologies are a heterogeneous group of metabolic perturbations characterized by morphological abnormalities and/or OXPHOS dysfunction [100]. Mitochondrial DNA analysis has revealed specific mutations for some mitochondriopathies. Although the specific OXPHOS mutations causing the disease may appear in all tissues, the functioning of only some of them is altered. The organ's sensitivity might be related to the different flux control coefficients of the mutated enzyme in the different tissues (Table 3) and to their ATP supply dependence from OXPHOS versus glycolysis.

MCA allows for the analysis of a metabolic flux or intermediate concentration by focusing either on one step or by grouping enzymes in blocks or in pathways. Thus, a comparative analysis of OXPHOS control distribution reveals that heart, kidney, some fast growing tumors (rat AS-30D hepatoma, mouse fibrosarcoma, human breast, lung, thyroid carcinoma, melanoma) [101], and perhaps skeletal muscle are more susceptible to mitochondrial mutations in ATP synthase, which is the only PS site with subunits encoded in the mitochondrial genome. On the other side, liver and brain might be more susceptible to mitochondrial mutations of the respiratory chain enzymes (see Table 3). Considering that the brain is a fully aerobic organ [102], whereas the liver depends on both OXPHOS (70–80%) and glycolysis (20–30%) for ATP supply [103], then it can be postulated that the brain is more sensitive to mutations in the mitochondrial genome than the liver because subunits of complexes I, III, and IV are encoded by the mitochondrial genome.

Titration of flux with specific inhibitors to determine the flux control coefficients of OXPHOS has been applied to intact tumor cells [90]. The results showed that the flux control resided mainly in site 1 of the respiratory chain ($C_{\text{Sitel}}^{\text{JOxPhos}} = 0.30$), whereas the other evaluated sites exerted a marginal control [90]. This observation could have therapeutic application if site 1 does not exert control in healthy cells, leading to less severe side effects.

The use of inhibitors in intact cells to determine control coefficients might pose two problems: hydrophilic inhibitors such as carboxyatractyloside (for ANT) and $\alpha$-cyano-4-hydroxy-cinammate (for pyruvate transporter) cannot readily enter the cell due to the presence of the plasma membrane barrier; the other problem is that hydrophobic but slow inhibitors, such as oligomycin, require long incubation times to ensure the interaction with the specific sites. These problems can be solved by incubating the cells for long periods of time and taking care of cell viability, for instance, AS-30D hepatoma cells are fairly resistant to this mechanical manipulation as they maintain high viability after a lengthy incubation under smooth orbital agitation of 1 h at 37°C [90].

### 5.3. Elasticity analysis

MCA defines the elasticity coefficients as

$$\varepsilon_X^{v_i} = \frac{dvi}{dX} \bullet \frac{X_o}{v_{io}}, \tag{6}$$

which is a dimensionless number that show the rate variation $v$ of a given enzyme or transporter $i$ when the concentration of a ligand $X$ (substrate $S$, product $P$ or allosteric modulator) is varied in infinitesimal proportions. The elasticity coefficients are positive for those metabolites that increase the enzyme or transporter rate (substrate or activator), and they are negative for the metabolites that decrease the enzyme or transporter rates (product or inhibitor). An enzyme working, under a steady-state metabolic flux, at saturating conditions of $S$ or $P$, is no longer sensitive to changes in these metabolites. Thus, its elasticity is close to zero (Figure 6, $\varepsilon_X^{v_i} = 0$). In turn, an enzyme working at $S$ or $P$ concentrations well below the Michaelis constant ($Km_S$ or $Km_P$) is expected to be highly sensitive to small variations in these metabolites (Figure 6, $\varepsilon_X^{v_i} = 1$).

The elasticities are intrinsically linked to the actual enzyme kinetics. If the kinetic parameters of an enzyme are known ($Vm_f$, $Vm_r$, $Km_S$, and $Km_P$), then the enzyme elasticity for any given metabolite concentration may be calculated as shown in the following equations.

For substrate,

$$\varepsilon_s^{v_i} = \frac{-S/Km_s}{1 + S/Km_s + P/Km_p} + \frac{1}{1 - \Gamma/\text{Keq}}, \tag{7}$$

and for product,

$$\varepsilon_P^{v_i} = \frac{-P/Km_P}{1 + S/Km_s + P/Km_P} - \frac{\Gamma/\text{Keq}}{1 - \Gamma/\text{Keq}}, \tag{8}$$

in which $\Gamma$ is the mass action ratio, and Keq is the equilibrium constant preferentially determined under physiological conditions.

An enzyme with low elasticity cannot increase (or decrease) its rate despite large variations in $S$ (or $P$) concentration; in consequence, such enzyme exerts a high flux control. In turn, an enzyme with a high elasticity can adjust its rate to the variation in $S$ or $P$ concentrations, and thus it does not interfere with the metabolic flux, exerting a low flux control. This inverse relationship between the elasticity and the flux control coefficients is expressed in a formal equation denominated connectivity theorem. A metabolic pathway can be divided in two blocks around an intermediary $X$: the producing (synthetic, supply) and the consuming (demand) enzyme blocks of $X$ are $i_1$ and $i_2$, respectively. Thus, the connectivity theorem for this two-block system is

$$\frac{C_{v1}^J}{C_{v2}^J} = -\frac{\varepsilon_X^{v2}}{\varepsilon_X^{v1}}. \tag{9}$$

The negative sign of the right part of the equation cancels with $\varepsilon_X^{vi1}$, which is negative because $X$ is a product of enzyme block $i_1$ (Figure 6).

To obtain the flux control coefficients, this approach requires experimental determination of the elasticity coefficients. How can this be done? Many strategies have been designed [90, 103–108], but the most used and probably more trustworthy is that in which the initial pathway metabolite ($S_o$) concentration is varied to increase the $X$ concentration (any intermediary in the pathway), and measuring in parallel the variation in flux. Under steady-state conditions, the flux rate is equal to the rate of end-product formation (i.e., lactate or alcohol for glycolysis; oxygen consumption for OXPHOS) and to the rate of any partial reaction. Then, plots of $X$ versus flux (Figure 7) are generated. The slope, calculated at the reference coordinate

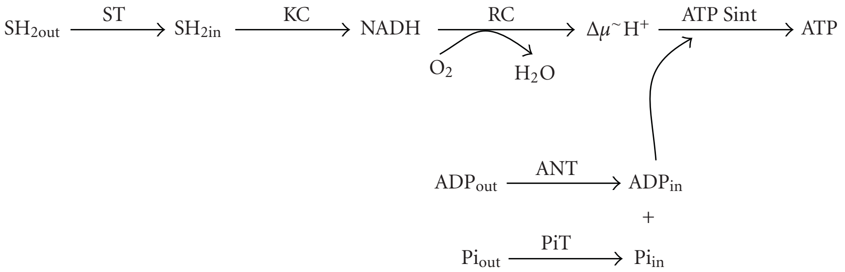
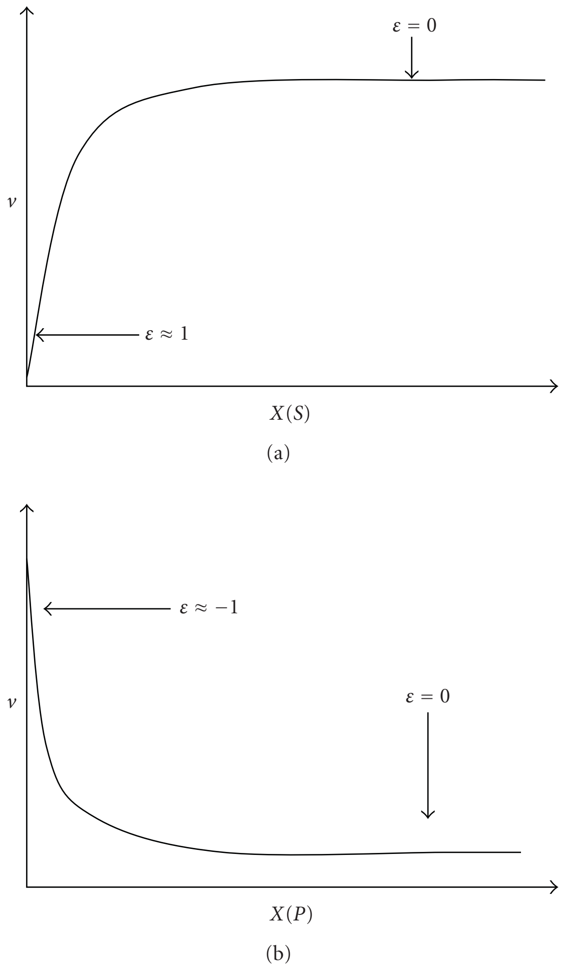


FIGURE 5: Mitochondrial oxidative phosphorylation. ST, oxidizable substrate transporter; KC, Krebs cycle; RC, respiratory chain; ($\Delta\tilde{\mu}_{H}{}^{+}$), proton electrochemical gradient; ANT, adenine nucleotide translocator; PiT, phosphate transporter; ATP Sint, ATP synthase.



(a)

(b)

FIGURE 6: Elasticity coefficients.

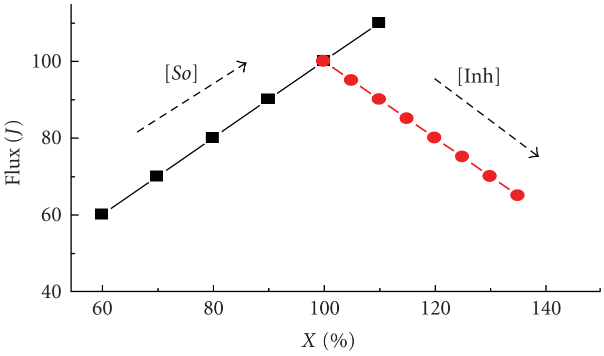


FIGURE 7: Experimental determination of the elasticity coefficients for substrates and products.

$(X_o, J_o)$ that is equivalent to $(S_o, v_{io})$, yields the elasticity coefficient of the consuming block of $X$. In another set of experiments, an inhibitor is added to block one or more enzymes after $X$. The $X$ concentration and flux are determined and plotted as shown in Figure 7, from which the elasticity coefficient of the producing block is calculated.

The flux control coefficients are determined by using the connectivity theorem and considering that the sum of the control coefficients comes to 1, $C_1 + C_2 = 1$ (summation theorem):

$$C^J_{v1} = \frac{\varepsilon^{v2}_X}{\varepsilon^{v2}_X - \varepsilon^{v1}_X}, \qquad C^J_{v2} = -\frac{\varepsilon^{v1}_X}{\varepsilon^{v2}_X - \varepsilon^{v1}_X}. \tag{10}$$

This method for determining $C^J_{v_i}$ using the elasticities of the two blocks was called double modulation by Kacser and Burns [83]. Years later, Brand and his group [103, 104] renamed this method as top-down approach. By applying the procedure shown in Figure 7 and using (10) for different metabolites along the metabolic pathway, it is possible to identify those sites that exert a higher control (which may be the sites for therapeutic use or biotechnological manipulation) and those that exert a negligible control under a given physiological or pathological situation.

Elasticity analysis has been used to evaluate the OXPHOS control distribution in tumor cells [90]. Almost all studies on this subject have been carried out with isolated mitochondria incubated in sucrose-based medium at 25 or 30°C or with the more physiological KCl-based medium but still at 30°C (Table 3). Furthermore, these studies did not consider that the product, ATP, never accumulates in the living cells, which does occur in experiments with isolated mitochondria. Under such a condition, a steady state in ATP production can never be reached as in living cells. In other words, the distribution of control in mitochondria (Table 3) has been determined in the absence of an ATP-consuming system. A remarkable exception to this incomplete experimental design was the work done by Wanders et al. [105], in which isolated liver mitochondria were incubated with two different ATP-consuming systems (or ADP-regenerating systems): HK + glucose and creatine kinase (CK) + creatine. Under this more physiological setting, the OXPHOS flux control distributed between ANT and the ATP-consuming system; however, flux control by the other pathway components was not examined. Therefore, to accurately evaluate OXPHOS control distribution, mitochondria should be incubated in the presence of an ATP-consuming system or in their natural environment (i.e., inside the cell).

The rate of OXPHOS in intact cells is determined from the rate of oligomycin-sensitive respiration: in the steady state, the enzyme rates are the same and constant; in branched pathways the sum of the branched fluxes equals the flux that supplies the branches. The global elasticity of the ATP-consuming processes (e.g., synthesis of protein, nucleic acid, and other biomolecules, as well as ion ATPases to maintain the ionic gradients, mechanical activity such as muscular contraction or flagellum and cilium movement, and secretion of hormones, digestive enzymes and neurotransmitters) is estimated by inhibiting flux with low concentrations of oligomycin or a respiratory chain inhibitor. To determine the elasticity of the ATP-producing block, flux, and [ATP] are varied with streptomycin, an inhibitor of protein synthesis (Figure 7). The elasticity coefficients are calculated from the initial coordinate slopes (without inhibitors) of each titration. With this procedure, it has been determined that the ATP-consuming block exerts a significant flux control of 34% [90]. Remarkably, this flux control value obtained in cells is quite similar to the flux control coefficients of the ATP-consuming system (HK or CK) reported by Wanders et al. [105] with isolated mitochondria.

Elasticity analysis by enzyme blocks allows the inclusion of the end-product demand as another pathway block. The conclusions obtained from this analysis have formulated the supply-demand theory [30], which proposes that when flux is controlled by one block (demand), the concentration of the end-product is determined by the other block (supply). The ratio of elasticities determines the distribution of flux control between supply and demand blocks. For instance, if $\varepsilon_X^{\text{Supply}} > \varepsilon_X^{\text{Demand}}$ (i.e., demand becomes saturated by the end-product $X$, and hence its elasticity is near zero), then the demand block exerts the main flux control. For concentration control, at larger $\varepsilon_X^{\text{Demand}} - \varepsilon_X^{\text{Supply}}$, smaller absolute values of both $C_{\text{Supply}}^X$ and $C_{\text{Demand}}^X$ are attained; hence, under demand saturation, the supply elasticity fully governs the magnitude of the variation in the end-product concentration. On the other hand, when demand increases, it loses flux control and induces a diminution in the end-product concentration. In turn, supply gains flux control and loses concentration control. In the presence of feedback inhibition, the system can maintain the end-product concentration orders of magnitude away from equilibrium (at a concentration around the $K_{0.5}$ of the allosteric enzyme).

As mentioned before, the demand is not usually included in the pathway because it is erroneously thought that it is not part of it. But then, is it valid to analyze the control of a metabolite synthesis if its demand is not considered? When the demand block is not included, it is assumed that the metabolic pathway produces a metabolite at the same rate regardless whether the metabolite demand is high or low. This reasoning is incorrect because a metabolic pathway indeed responds to changes in the metabolite demand and, more importantly, a pathway without end-products consumption reactions is unable to reach a steady state.

Therefore, a metabolic pathway can be divided in supply and demand blocks. The intermediary $X$ linking the two blocks is one of the end-products of the producing block (e.g., pyruvate or lactate or ethanol, and ATP for glycolysis). The variation in rate of the two blocks in response to a variation in $X$ can be theoretical or experimentally determined (Figure 8(a)). It is worth noting that, for this supply-demand approach, it is not necessary to know the kinetics of each pathway enzyme because the rate response of each block reflects the global kinetics of all participating enzymes. When the $X$ concentration is increased, the rate of the supply block decreases (i) because $X$ is its product and (ii) because usually an enzyme within this block receives information from the final part of the pathway, decreasing its rate through feedback inhibition. In turn, the rate of the demand block increases as $X$ is its substrate.

To better visualize the effect of large rate changes, the kinetics of both blocks are plotted in a logarithmic scale. Figure 8(b) shows the kinetics described in Figure 8(a) converted to natural logarithm. The intersection point between kinetic curves, at which the supply and demand rates are identical, represents the pathway steady-state flux (in the $Y$ axis) and end-product concentration (in the $X$ axis). Since the elasticity is also defined as $\varepsilon_X^{v_i} = d \ln v_i / d \ln X$, the slope at the intersection point represents the elasticity of each block towards the intermediary $X$. Here, the use of the scalar factor is not necessary because it is included in the logarithmic equation. With the elasticity coefficients calculated from plots like those shown in Figure 8, and the connectivity theorem, the flux control coefficient of each block is determined. The example in Figure 8(b) shows that the demand exerts a high flux control (and has low elasticity) and the supply block exerts low control (and has high elasticity).

The fact that the demand may exert a high flux control in metabolite pathways has at least three important implications: (a) the supply block responds to variations in the demand (high elasticity); (b) the demand block has information transfer mechanisms towards the supply block that avoid the unrestricted intermediary accumulation under a low demand, particularly when the supply block has reactions with large Keq (>100; $\Delta G^{\circ\prime} > 3\,\text{Kcal}\,\text{mol}^{-1}$ at 37°C); and (c) if the main flux control resides in the demand block, then the supply block may only exert control on the intermediary concentration but not on the flux [30, 32]. This last conclusion explains why it is incorrect to consider that an enzyme that controls flux must also control the intermediary concentration.

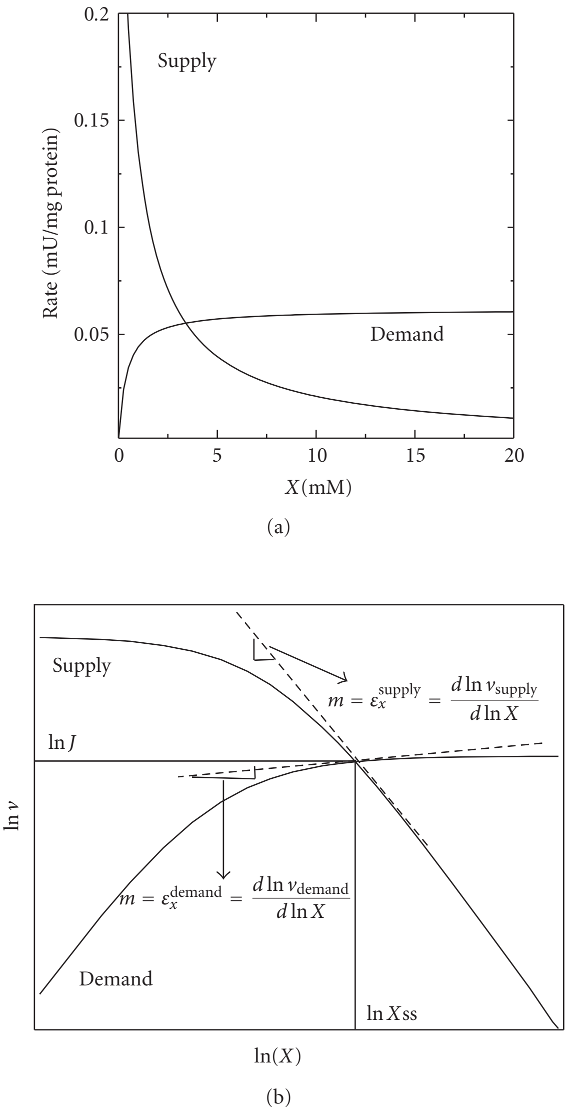


Figure 8: (a) Kinetics of the synthesis (supply) and consuming (demand) blocks of the intermediary $X$. The kinetic parameters are from enzymes in tobacco glutathione (GSH) synthesis. $X$ represents the intermediary concentration, in this case GSH. (b) Rate plots of the supply and demand blocks in a natural logarithmic scale.

Regulatory mechanisms of enzyme activity are modulation of protein concentration by synthesis and degradation, as well as covalent modification and variation in the substrate or product concentrations (which are components of the pathway). In addition, another regulatory mechanism is the modulation by molecules that are not part of the pathway, that is, through allosteric interaction with cooperative (sigmoidal kinetics) or noncooperative enzymes (hyperbolic kinetics) (e.g., $Ca^{2+}$ activates some Krebs cycle dehydrogenases; citrate inhibits PFK-1; malonyl-CoA inhibits the mitochondrial transporter of acyl-carnitine/carnitine; or the initial substrate of a pathway that has not entered the system). For these last cases, Kacser and Burns [83] proposed the use of the response coefficient $R$ which is defined by the following expression:

$$R_M^J = C_{v_i}^J \bullet \varepsilon_M^{v_i}, \qquad (11)$$

where $M$ is the external modulator of the $i$ enzyme. The response coefficient is $dJ/dM \bullet M_o/J_o$. If the elasticity of the sensitive enzyme toward the external effector is also determined, then it is possible to calculate $C_{v_i}^J$ by using (11). Unfortunately, due to the experimental complexity for determining the elasticity coefficient, this coefficient is often calculated in a theoretical way by using the respective rate equation (Michaelis-Menten or Hill equations) and the kinetic parameters $Km$ and $V_{\max}$ determined by someone else under optimal assay conditions, which are commonly far away from the physiological ones. Therefore, for this theoretical determination of elasticity only the value of the external modulator concentration is required. It is convenient to emphasize that the determination of the flux control coefficients becomes more reliable when they are calculated from several experimental points (Figure 7), instead of only one, as occurs with the theoretical elasticity analysis.

Groen et al. [106] determined the flux control distribution of gluconeogenesis from lactate in hepatocytes by using both theoretical and experimental elasticity analysis and the response coefficient. These authors concluded that gluconeogenesis stimulated by glucagon was controlled by the pyruvate carboxylase ($C_{\text{PC}}^{J_{\text{glucose}}} = 0.83$); in the absence of this hormone, the control was shared by PC, PYK, ENO-PGK segment, and TPI-fructose-1,6-biphosphatase segment [106].

Elasticity analysis has been applied to elucidate the flux control of ATP-producing pathways in fast-growing tumor cells. For OXPHOS, this approach showed that respiratory chain complex I and the ATP-consuming pathways were the enzymes with higher control ($C_{v_i}^J = 0.7$) [90]. For glycolysis, the main flux control ($C_{v_i}^J = 0.71$) resided in GLUT + HK reactions because HK is strongly inhibited by its product G6P despite extensive enzyme overexpression [107]. Examples of elasticity analysis on other pathways are photosynthesis [108], ketogenesis [109], serine [110] and threonine synthesis in *E. coli* [111], glycolysis in yeast [112], glucose transport in yeast [113], DNA supercoiling [114], glycogen synthesis in muscle [115], and galactose synthesis in yeast [116].

In conclusion, the elasticity analysis is the most frequently used method for determining flux control coefficients because it does not need a group of specific inhibitors for all the enzymes and transporters of the pathway, neither does it require knowledge of the inhibitory mechanisms or kinetic constants. It is only necessary to produce a variation in the intermediary concentration $X$ by using an inhibitor of either block or by directly varying the $X$ concentration.

### *5.4. Pathway modeling*

In agreement with Fell [2], it seems impossible for a researcher to analyze one by one the rate equation of each enzyme in a metabolic pathway to predict and explain the system behavior as a whole. To deal with this problem, in the last three decades some scientists have constructed mathematical models for some metabolic pathways using several software programs. Thus, the specific variation of a single enzyme activity without altering the rest of the pathway (Figure 4), which has been an experimentally difficult task for applying MCA, becomes easier to achieve with reliable computing models. The term "in silico biology" has been coined for this approach.

There are two basic types of modeling: (a) structural modeling and (b) kinetic modeling. The former is related to the pathway chemical reaction structure and does not involve kinetic information. The use of reactions is based on their stoichiometries. The information obtained with structural modeling is the description of the following:

(i) the exact determination of which reactions and metabolites interact among them;

(ii) the conservation reactions. There are metabolites for which their sum is always constant or conserved (e.g., NADH + $NAD^+$; NADPH + $NADP^+$; ubiquinol + ubiquinone; ATP + ADP + AMP; CoA + acetyl-CoA). The identification of conserved metabolites might not be obvious;

(iii) enzyme groups catalyzing reactions in a given relationship with another group of enzymes;

(iv) elemental modules, which are defined as the minimal number of enzymes required to reach a steady state, which can be isolated from the system (for a review about structural modeling; see [117]).

Kinetic modeling is more frequently used. In addition to an appropriate computing program, this approach requires the knowledge of the stoichiometries, rate equations, and Keq values of each reaction in the pathway (or the $V_{\max}$ in the forward and reverse reactions), as well as the intermediary concentrations reached under a given steady state. Some currently used softwares are Copasi (http://www.copasi.org/tiki-index.php) based on Gepasi (http://www.gepasi.org/; [118]); Metamodel [119]; WinScamp [120] and Jarnac [121] (both available at http://www.sys-bio.org/); and PySCeS (http://pysces.sourcesforge.net/; [122]). For other programs and links, go to http://sbml.org/index.psp. To reach a steady-state flux, it is necessary to fix the initial metabolite concentration to a constant value and the irreversible and constant removal of the end products. Except for the final reactions in which their products have to be removed from the system, all pathway reactions have to be considered as reversible, notwithstanding whether they have large Keq (if there is an irreversible reaction under physiological conditions, then a reversible rate equation that includes the Keq suffices to maintain the reaction as practically irreversible). Care should be taken to include the enzyme's sensitivity toward its products because this property is related with the enzyme elasticity and hence with its flux control; omission of this parameter may very likely lead to erroneous conclusions.

It should be pointed out that the purpose of kinetic modeling is not merely to replicate experimental data but also to explain them [117]. Thus, pathway modeling is a powerful tool that allows for (i) the detection of those properties of the pathway that are not so obvious to visualize when the individual kinetic characteristics of the participating enzymes are examined; and (ii) the understanding of the biochemical mechanisms involved in flux and intermediary concentration control. Modeling requires the consideration of all reported experimental data and interactions that have been described for the components of a specific pathway, thus allowing for the integration of disperse data, discarding irrelevant facts [84]. Although all models are oversimplifications of complex cellular processes, they are useful for the deduction of essential relationships, for the design of experimental strategies that evaluate the control of a metabolic pathway, and for the detection of incompatibilities in the kinetic parameters of the participating enzymes and transporters, which may prompt the experimental revision of the most critical uncertainties.

With the model initially constructed, the simulation results do not usually concur with the experimental results; in consequence, the model normally requires refinement, a point at which the researcher's thinking and knowledge of biology plays a fundamental role in modifying the structure and parameters of the model. The discrepancies observed between modeling and experimentation unequivocally pinpoint what elements or factors have to be re-evaluated or incorporated so that the model approximates more closely reality (i.e., experimental data). The comparison of the experimentally obtained intermediary concentrations and fluxes with those obtained by simulation is an appropriate validating index of the model; this index indicates whether the model approximation to the physiological situation is acceptable or whether re-evaluation of the kinetic properties of some enzymes and transporters and/or incorporation of other reactions or factors is required.

A reason to why the results obtained by modeling may substantially differ from the experimental results is that the kinetic parameters of the pathway enzyme and transporters and the Keq values used were determined by different research groups, under different experimental conditions and in different cell types. Moreover, enzyme kinetic assays are carried out at low, diluted enzyme concentrations (thus discarding or ignoring relevant protein-protein interactions), and at optimal (but not physiological) pH and "room temperature" (which may be far away from the physiological values). In addition, no experimental information is usually available regarding the reactions reversibility and the product inhibition of the enzymes and transporters (particularly for physiological irreversible reactions, i.e., reactions with large Keq). With worrisome frequency, the researcher has to adjust the experimentally determined $Vm$ and $Km$ values to achieve a model behavior that acceptably resembles that observed in the biological system. Apparently, this type of limitations as well as the sometimes overwhelming amount of kinetic data necessary for the construction of a kinetic model has restricted the number of reliable models that can be used for the prediction of the pathway control structure.

Once the kinetic model stability, robustness, structural and dynamic properties have been evaluated, and experimentally validated, the model may become a virtual laboratory in which any parameter or component can be modified or replaced and any aspect of the pathway behavior can be explored within a wide diversity of circumstances or limits [117]. At this stage, the model is suitable for examining the pathway regulatory properties and control structure.

Glycolysis in *S. bayanus, S. cerevisiae* [113, 123, 124], and *Trypanosoma brucei* [125, 126] is the metabolic pathway that has been more extensively modeled. Both cell types have a very active glycolysis and are fully dependent on this metabolic pathway for ATP supply, under anaerobiosis and aerobiosis, respectively. One advantage of modeling glycolysis in these cell types is that most of the kinetic parameters used have been experimentally determined by the same groups under the same experimental conditions. However, the kinetics of the reverse reactions has not been determined and thus these authors used $Km_P$ and Keq values reported by others and obtained in other cell types under rather different experimental conditions, or they were adjusted to improve model fitting.

Nevertheless, the simulation results yielded relevant information on the control of the glycolytic flux. In both cases, the enzymes traditionally considered the rate-limiting steps, HK, ATP-PFK-1, and PYK did not contribute to the flux control, whereas the main control resided in GLUT (54% in the parasite and 85–100% in yeast). Under some conditions, HK may exert some control (15%) in *S. cerevisiae* and some nonallosteric enzymes such as ALDO, GAPDH, and PGK may also exert some flux control in *T. brucei*.

MCA through kinetic modeling has been applied to several pathways:

(i) glycolysis in erythrocytes [84] in which flux control distributes between HK (71%) and PFK-1 (29%);

(ii) carbohydrate metabolism during differentiation in *Dictyostelium discoideum* [127] with cellulose synthase (86%) as the main controlling step;

(iii) sucrose accumulation in sugar cane with HK, invertase, fructose uptake, glucose uptake, and vacuolar sucrose transporter having the most significant flux control [128];

(iv) glycerol synthesis in *S. cerevisiae* with GAPDH (85%) as the main control step [129];

(v) penicillin synthesis in *Penicillium chrysogenum* controlled (75–98%) either by d-(a-aminoadipyl) cysteinylvaline synthetase (short incubation times <30 hour) or isopenicillin N. synthetase (long incubation times > 100 h) [130];

(vi) Calvin cycle [131] controlled by GAPDH (50%) and sedoheptulose-1,7-bisphosphatase (50%);

(vii) threonine synthesis in *E. coli* controlled by homoserine dehydrogenase (46%), aspartate kinase (28%), and aspartate semialdehyde dehydrogenase (25%) [111];

(viii) lysine production in *Corynebacterium glutamicum* mainly controlled by aspartate kinase and permease [132];

(ix) nonoxidative pentose pathway in erythrocytes mainly controlled by transketolase (74%) [133];

(x) EGF-induced MAPK signaling in tumor cells controlled by Ras-activation by EGF (21%), Ras dephosphorylation (43%), ERK phosphorylation by MEK (44%), and MEK phosphorylation by RAS (143%) [13];

(xi) *Aspergillus niger* arabinose utilization with flux control shared by arabinose reductase (68%), arabitol dehydrogenase (17%), and xylulose reductase (14%) [134];

(xii) glycolysis in *L. lactis* in which several end products are generated (lactate, organic acids, ethanol, acetoin) [135]. Model predictions indicated that flux toward diacetyl and acetoin (important flavor compounds) was mainly controlled by LDH but not by acetolactate synthetase, the first enzyme of this branch.

We modeled the GSH and PCs biosynthesis (Figure 2) to determine and understand the control structure of the pathway and thus be able to identify potential sites for genetic engineering manipulation that might lead to the generation of improved species in heavy metal resistance and accumulation. Two models were constructed, one for higher plants and the other for yeast, both exposed to high concentrations of $Cd^{2+}$ [136]. Due to the similarity in the results, only the plant results are analyzed below.

An interesting conclusion from the GSH-PCs synthesis modeling is that control of flux (and GSH concentration) is shared between the GSH supply and demand under both unstressed and $Cd^{2+}$ exposure conditions (Table 4). This observation strongly differs from the idea that $\gamma$-ECS is the rate-limiting step [33–35]. For many researchers, the concept of $\gamma$-ECS being the key controlling step has seemed to be correct because (a) $\gamma$-ECS receives information from the final part of the pathway, as it is potently inhibited by GSH, the pathway end-product; and (b) $\gamma$-ECS is localized in the first part of the pathway (Figure 2). In addition, GS is usually more abundant and efficient than $\gamma$-ECS [137].

However, in most of the studies on the control of GSH synthesis, the GSH demand has not been considered. The GSH synthesis modeling shows that under a physiological feedback inhibition of $\gamma$-ECS by GSH a small increase in demand increases flux because the GSH concentration decreases and the $\gamma$-ECS inhibition attenuates. In contrast, if the demand remains constant, then an increase in $\gamma$-ECS activity or content (by overexpression) does not increase flux because the GSH inhibition is still there and operates on both new and old enzymes. The same pattern is also observed when HK is overexpressed to increase glycolytic flux since it is still inhibited by G6P (see Section 3). On the other hand, $\gamma$-ECS indeed exerts significant concentration control on GSH, which means that a $\gamma$-ECS increase results in higher GSH concentration (Table 4). This last observation demonstrates

Table 4: Control of GSH and PC synthesis in plants exposed to $Cd^{2+}$.

| Enzyme | 1x $\gamma$-ECS + PCS | | | | 2.5x $\gamma$-ECS + PCS | | | |
|---|---|---|---|---|---|---|---|---|
| | $C_{vi}^{JGSH}$ | $C_{vi}^{JPC}$ | $C_{vi}^{GSH}$ | $C_{vi}^{PC}$ | $C_{vi}^{JGSH}$ | $C_{vi}^{JPC}$ | $C_{vi}^{GSH}$ | $C_{vi}^{PC}$ |
| $\gamma$-ECS | 0.58 | 0.60 | 0.68 | 0.76 | 0.45 | 0.61 | 0.70 | 0.60 |
| GS | <0.01 | <0.01 | 0.01 | 0.01 | 0.19 | <0.01 | <0.01 | 0.97 |
| GS-transferase | 0.01 | −0.06 | −0.07 | −0.07 | <0.01 | <0.01 | < −0.01 | −0.05 |
| PCS | 0.40 | 0.44 | −0.63 | −0.56 | 0.33 | 0.44 | −0.62 | 0.57 |
| vacuole PC-Cd transporter | <0.01 | <0.01 | <0.01 | −1.2 | <0.01 | <0.01 | <0.01 | −2.1 |

$C_{vi}^{JGSH}$, control coefficient of enzyme *i* in GSH synthesis; $C_{vi}^{JPC}$, control coefficient of enzyme Ei on PCs synthesis; $C_{vi}^{GSH}$, control coefficient of enzyme *i* on GSH concentration; $C_{vi}^{PC}$, control coefficient of enzyme *i* on PCs concentration. An enzyme with a negative flux control indicates that it is localized in a branch, turning aside the principal flux; an enzyme with a negative concentration control indicates that an increase in its activity decreases metabolite concentration.

that an enzyme controlling a metabolite concentration does not necessarily control the flux.

$Cd^{2+}$ exposure promotes a high GSH demand because significant oxidative stress surges, thus causing oxidation of GSH through GSH peroxidases, and because GSH and PCs are used for sequestering the toxic metal ion; hence, a higher GSH consuming rate sets up. Under this condition, modeling predicted that control was almost equally shared between the supply and demand blocks, but particularly between $\gamma$-ECS and PCS (see Figure 2). Modeling was also able to explain why PCS overexpression can have toxic effects on the cell [36]. An increase in the GSH demand (PCS overexpression) under high-demand conditions ($Cd^{2+}$ stress) leads to GSH depletion that severely compromises other processes such as the oxidative stress control and xenobiotic detoxification.

The conclusions drawn by this model led us to propose that, to significantly increase the $Cd^{2+}$ resistance and accumulation, $\gamma$-ECS and PCS should be simultaneously overexpressed (Table 4; Figure 9). This particular manipulation promotes an increase in the rate of GSH and PCs synthesis (determined by the high-to-low transition of their flux control coefficients) and in the GSH and PCs concentrations (determined by their high concentration control coefficients). The model predicts that a 2-fold increase in the simultaneous overexpression of $\gamma$-ECS and PCS brings about a 1.9–2.4-fold increase in flux to GSH ($J_{GS}$) and PCs ($J_{PCS}$) and in PCs concentration (Figure 9); a 5-fold overexpression further increases by 4.5–8.1 times the fluxes and PCs concentration.

This proposed enzyme overexpression should not exceed the GS and the complex PC-Cd (or GS-Cd-GS) vacuolar transporters' maximal activities, in order to keep the cell away from a severe oxidative stress caused by GSH depletion or $\gamma$-EC accumulation. Indeed, the concentration of GSH was maintained high and constant although $\gamma$-EC accumulated with the simultaneous overexpression (Figure 9). Furthermore, this enzyme manipulation should avoid the increase of the PC-Cd and GS-Cd-GS complexes in cytosol to toxic levels. In other words, excessive enzyme overexpression should be avoided, unless this is accompanied by compensating overexpression of consuming enzymes (GS for $\gamma$-ECS overexpression and PCs vacuolar transporters for that of PCS). In yeasts and plants, $Cd^{2+}$ is ultimately inactivated by the additional interaction with $S^{2-}$ and the subsequent formation of stable high molecular weight complexes with PCs, $Cd^{2+}$, $S^{2-}$, and GSH [138, 139]. In parallel to the $\gamma$-ECS and PCS overexpression, moderate repression of GSH-S-transferases, which compete for the available GSH (Figure 2), may also promote an increase in GSH concentration and PCs formation flux [136].

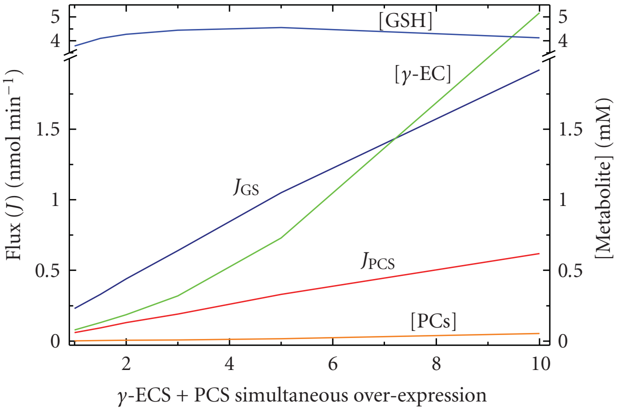


Figure 9: Modeled simultaneous overexpression of two controlling enzymes, one in the supply ($\gamma$-glutamylcisteine synthetase, $\gamma$-ECS) and the other in the demand branch (phytochelatin synthase, PCS), of the glutathione and phytochelatins synthesis pathway in plants.

MCA is based on infinitesimal changes in an enzyme or metabolite concentration. In contrast, gene overexpression induces large changes in activity; hence, further theoretical background has been developed for predicting the effect on flux and metabolite concentrations induced by large enzyme changes. Such a theoretical background was initially developed by Small and Kacser [140], who depicted (12) based on the flux control coefficients to predict the effect promoted by large changes in enzyme activity:

$$f^{J}_{E^{r}_{j-m}} = \frac{1}{1-\sum_{i=j}^{m}\left(C^{J_o}_{vi0}\bullet(r_i-1)/r_i\right)}, \qquad (12)$$

in which $f$ is the amplification factor (the flux increase), and $r$ represents how many times the enzyme is overexpressed. To predict the flux changes, promoted by identical

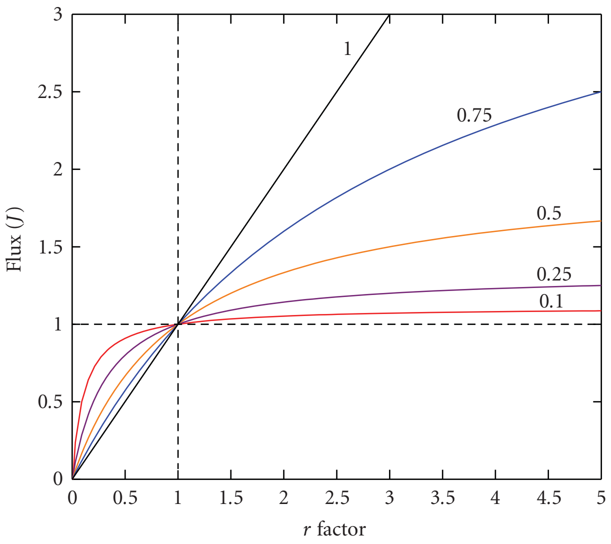


Figure 10: Effect on flux when one or more enzymatic activities with different control coefficients are varied. This figure represents an enzyme or group of enzymes in which their $C^J_{v_i}$ sum is indicated in parenthesis and is modified by the same $r$ factor. Number 1 represents the reference control, thus if $r < 1$, there is suppression, whereas $r > 1$ represents overexpression.

overexpression of two enzymes (same $r$ value) with different $C^J_{v_i}$, the equation is

$$f^J_{E^r_{j-m}} = \frac{1}{1 - (C^J_i + C^J_j) \bullet ((r-1)/r)}. \qquad (13)$$

Figure 10 shows the effect on flux when one or more enzymes with different $C^J_{v_i}$ are changed by the same $r$ factor. If the sum of $C^J_{v_i}$ of one or more enzymes is less than 0.25, the impact on flux is discrete when the expression increases 5 folds (which is the most common variation in the overexpression experiments analyzed in Section 2). But for a 3-fold overexpression of a group of enzymes, for which their sum of $C^J_{v_i}$ is more than 0.5, then a significant flux change is achieved. If the sum of $C^J_{v_i}$ is 1, the flux varies in a linear proportion with the degree of overexpression. It has to be remarked, however, that the predicted change in flux (Figure 10) will be valid until certain degree, the limits of which being determined by the other pathway enzymes that should stay as noncontrolling steps.

Figure 10 also shows the effect on flux of decreasing an enzyme activity (third quadrant). This segment plot is useful when inhibition of pathway flux is being pursued for therapeutic purposes or for understanding the molecular basis of the genetic dominance and recessivity. Like in the enzyme overexpression experiment, only a significant effect on flux is achieved when the enzymes with high $C^J_{v_i}$ values are inhibited. For an enzyme or group of enzymes with $C^J_{v_i}$ of 0.25, greater than 80% inhibition has to be attained to decrease 50% the pathway flux. In this context, it seems feasible to explain why knockdown of enzymes involved in $TSH_2$ synthesis has to be almost total to detect an effect on $TSH_2$ content or to alter functional or pathogenic properties of the parasites (Section 4.3). The knockdown or knockout experiments in trypanosomatids suggest that γ-ECS, TryS, and TryR most probably have low flux control and concentration-control coefficients since their contents or activities have to be reduced >80% of the normal levels to reach changes in intermediary levels or in oxidative stress handling.

Contrary to the several unsuccessful overexpression experiments carried out to increase the flux or metabolites of a metabolic pathway, modeling may allow for a more focused and appropriate design of experimental strategies of genetic engineering to increase flux or a given metabolite, and for selecting drug targets to decrease flux or metabolite concentration. For these predictions, modeling considers that overexpression of a controlling enzyme or transporter may promote flux or metabolite control redistributions. Thus, a low-control step may become a controlling point when overexpressing another step and, in consequence, the prediction shown in Figure 10 based on (11) and (12) may be inaccurate. By considering the whole pathway components, modeling is also a powerful tool for predicting the effects on flux and metabolite concentration of varying an enzyme activity (by overexpression or drug inhibition).

### *Model predictions to inhibit a pathway flux*

Kinetic modeling has been used to identify the flux controlling steps in *Trypanosoma brucei* glycolysis for drug targeting purposes. Interestingly, modeling has predicted controlling steps for the parasite pathway different from those described for glycolysis in human host cells [125, 126].

*Entamoeba histolytica* is the causal agent of human amebiasis. The parasite lacks functional mitochondria and has neither Krebs cycle nor OXPHOS enzyme activities. Therefore, substrate level phosphorylation by glycolysis is the only way to generate ATP for cellular work [141]. An important difference in amebal glycolysis in comparison to glycolysis in human cells is that it contains the pyrophosphate (PPi)-dependent enzymes phosphofructokinase (PPi-PFK) and pyruvate phosphate dikinase (PPDK), which replace the highly modulated ATP-PFK and PYK present in human cells. Moreover, both have been proposed as drug targets by using PPi analogues (bisphosphonates) [141].

We recently described the construction of a kinetic model of *E. histolytica* glycolysis to determine the control distribution of this energetically important pathway in the parasite [142]. The model was constructed using the Gepasi software and was based on the kinetic parameters determined in the purified recombinant enzymes [143], as well as the enzyme activities, fluxes, and metabolite concentrations found in the parasite. The results of the metabolic control analysis indicated that HK and PGAM are the main flux control steps of the pathway (73 and 65%, resp.) and perhaps GLUT. In contrast, the PPi-PFK and PPDK displayed low flux control (13 and 0.1%, resp.) because they have overcapacity over the glycolytic flux [142]. The amebal model allowed evaluating the effect on flux of "inhibiting" the pathway enzymes. The model predicted that

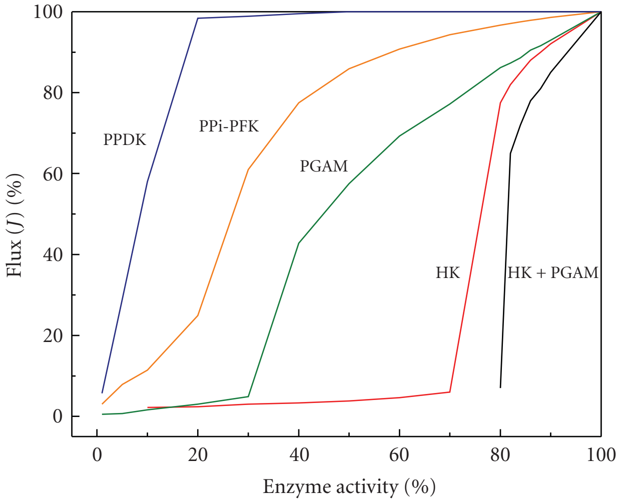


Figure 11: Modeled flux behavior when inhibiting pathway enzymes. The predicted flux when varying the enzyme activity was obtained using the kinetic model for *Entamoeba histolytica* glycolysis [142]. In this case, 100% enzyme activity is the enzyme activity present in amebal extracts, and 100% flux is the ethanol flux displayed by amoebae incubated with glucose. PPi-PFK, PPi-dependent phosphofructokinase; PPDK, pyruvate phosphate dikinase; PGAM, 2,3 bisphosphoglycerate independent 3-phosphoglycerate mutase.

in order to diminish by 50% the glycolytic flux (and the ATP concentration; data not shown), HK and PGAM should be inhibited by 24 and 55%, respectively, or both enzymes by 18% (Figure 11). In contrast, to attain the same reduction in flux by inhibiting PPi-PFK and PPDK, they should be decreased >70% (Figure 11). Therefore, the kinetic model results indicate that HK can be an appropriate drug target because its specific inhibition can compromise the energy levels in the parasite. They also indicate that although PPi-PFK and PPDK remain as promising drug targets because of their divergence from the human glycolytic enzymes, highly potent and very specific inhibitors should be designed for these enzymes in order to affect the parasite's energy metabolism.

### 5.5. *In vitro reconstitution of metabolic pathways*

Another experimental approach for determining the enzyme control coefficients is the in vitro reconstitution of segments of metabolic pathways. It is recalled that for determining the flux control coefficient exerted by a given step on a metabolic pathway the enzyme activity has to be varied, without altering the other components in the system, and the flux variations are to be measured (Figure 4). Such an experiment can be readily made if a pathway is reconstituted with purified enzymes. Some advantages of this approach are that the pathway structure is known, in which the components concentration may be manipulated and analyzed separately, and the enzyme effectors can be assayed. As the system composition is strictly controlled, the results may be highly reproducible. The main disadvantage is that the enzyme concentrations in the assays are diluted and thus the enzyme interactions are not favored. If this interaction is important for activity, the in vitro reconstitution may limit the extrapolation to the metabolic pathway inside the cell.

There are not many studies describing this type of experiments, most probably due to the fact that for applying MCA the pathway must be working under steady-state conditions. In a reconstituted system, only a quasi steady state may be reached because there is net substrate, and cofactors consumption, as well as product accumulation, since it is difficult to attain a constant substrate supply and release of products.

One of the first experimental reports on control coefficient determination in a reconstituted system was carried out for the upper glycolytic segment with the commercially available rabbit muscle HK, HPI, PFK-1, ALDO, and TPI [144]. Each enzyme was separately titrated and the flux variation to glycerol-3-phosphate (by coupling the reconstituted system to an excess of $\alpha$-GPDH) was measured in the presence of CK to maintain the ATP concentration constant. The flux control coefficients were determined as described in Figure 4. The results showed that PFK-1 and HK exerted the main flux control (65% and 20%, resp.), whereas the remaining 15% resided in the other enzymes. These authors observed that the addition of F1,6BP, a PFK-1 activator slightly diminished the flux control exerted by PFK-1 and increases that of HK. The validation of the summation theorem was also demonstrated in this work [144].

The lower glycolytic segment has also been reconstituted with commercial enzymes for determining the flux control coefficients [145]. The results showed that flux was mainly controlled by PYK (60–100%), although under some conditions control was shared with PGAM; ENO did not contribute to the flux control.

Another important limitation of the reconstitution experiments is that the commercial availability of the purified enzymes from the same organism is restricted or inexistent. However, by using the information from the genome sequence projects and the recombinant DNA technology, it is now possible to access all the enzyme genes from a metabolic pathway in the same organism, thus facilitating their cloning, overexpression, and purification. With this strategy, we cloned, overexpressed, and purified the 10 glycolytic enzymes of *Entamoeba histolytica* [143] for studying the flux control distribution in this organism by using kinetic modeling [142] and pathway reconstitution.

The reconstitution experiments of the lower amebal glycolytic segment, under near physiological conditions of pH, temperature, and enzyme activity (Figure 12) showed that PGAM and, to a lesser extent, PPDK exert the main flux control (these amebal enzymes are genetically and kinetically different from their human counterparts) with ENO exhibiting negligible control [143]. In turn, reconstitution of the upper amebal glycolytic segment has revealed that HK and, to a much lesser extent HPI, PPi-PFK, and ALD, exerted the main flux control, with TPI having negligible control [146]. These results strongly correlate with the enzyme catalytic efficiencies previously reported [143], in which HK is highly sensitive to AMP inhibition, ALD, and PGAM have the lowest catalytic efficiencies among the glycolytic enzymes, leading to high flux control coefficients and thus becoming suitable candidates for therapeutic intervention. The reconstitution results also agree with the pathway modeling predictions previously analyzed (Section 5.4), in which HK and PGAM are two of the main controlling steps [142].

The in vitro reconstitution experiments are also useful for studying the effect on control redistribution of an enzyme modulation that is particularly difficult to manage in vivo; the main controlling steps identified with the reconstitution experiments should be further analyzed with other experimental strategies such as elasticity analysis in the in vivo systems.

### 5.6. Genetic engineering to manipulate the in vivo protein levels

This experimental approach for determining the control coefficients could be part of the genetic approach analyzed in Section 5.1, but it was separated due to its recent methodological development and because it actually belongs to the molecular genetics rather than to the Mendelian genetics.

#### 5.6.1. Repression of gene expression

This approach is based on the in vivo modulation of the enzyme levels using the RNA antisense technology. There are at least three strategies to inhibit gene expression: (a) the use of single stranded antisense oligonucleotides, which form a double stranded RNA that might be degraded by RNAse H; (b) target RNA degradation with catalytically active oligonucleotides, known as ribozymes that bind to their specific RNA; and (c) RNA degradation using siRNAs (21–23 nucleotides) [147].

The RNA antisense technology was applied for control coefficient determination of the ribulose-bisphosphate-carboxylase (Rubisco) that fixes $CO_2$ in the plant Calvin cycle. This enzyme considered the rate-limiting step of the Calvin cycle and of the whole photosynthetic process, despite its high concentration (4 mM) in the chloroplasts stroma that compensates its low catalytic efficiency.

Attempts to make Rubisco a nonlimiting step, either by modifying its catalytic efficiency or by overexpressing it, have been unsuccessful. Stitt et al. [148] determined the $C^{J\text{photosynthesis}}_{\text{rubisco}}$ of tobacco plants by decreasing its activity with DNA antisense. The plants were transformed with DNA antisense against the mRNA of the enzyme's small subunit, thus promoting its degradation. For Calvin cycle enzymes, the pleiotropic effects were minimal. The results showed that Rubisco may indeed be the photosynthesis limiting step with a $C^{J\text{photosynthesis}}_{\text{rubisco}} = 0.69–0.83$ when plants are exposed to high illumination (1050 $\mu$mol quanta $m^{-2}s^{-1}$), high humidity (85%), and low $CO_2$ concentrations (25 Pa). However, this flux control decreases to 0.05–0.12 under moderate illumination or high $CO_2$ levels [148]. Unfortunately, the authors did not determine the control coefficients of the other pathway enzymes or the branches fluxes which may be significant.

As described in Section 5.4, the results of the *T. brucei* glycolysis modeling indicated that GLUT was the main flux control step ($C^J_{\text{GLUT}} \sim 50\%$), [125, 126]. This model predicted a large overcapacity for HK, PFK-1, ALDO, GAPDH, PGAM, ENO, and PYK over the glycolytic flux leading to low flux control coefficients [125, 126]. To validate the modeling results, the concentrations of HK, PFK-1, PGAM, ENO, and PYK were changed with siRNAs in growing parasites [149]. These knockdown expression experiments showed overcapacity of HK and PYK over the flux, although at lower levels than predicted by the model. A good correlation for PGAM and ENO was obtained between model predictions and experimental results. However, a large difference (9 folds) was obtained for PFK-1. This discrepancy is perhaps related to pleiotropic effects of PFK-1 downregulation, as these mutants also displayed diminution in the activities of other enzymes (HK, ENO, and PYK). The combination of these two approaches, in silico modeling and in vivo experimentation, is complementary: on one hand, modeling identifies the enzymes (out of 19 that contain the model) that display the highest flux control coefficients, whereas in vivo experimentation validates the accuracy of the model to establish predictions about the pathway's behavior.

#### 5.6.2. Fine tuning of cellular protein expression

The knockdown experiments described above usually yield only two experimental points of the plot shown in Figure 4: the wild-type and the knockdown strain protein levels or enzyme activities. Thus, with such an approach high levels of inhibition (>80%) are mostly analyzed, whereas intermediate levels of downregulation (if obtained) are generally overlooked. Therefore, knockdown experiments are not very useful to obtain the complete set of experimental data (above and below the wild-type levels of enzyme activity with the corresponding flux) for determining reliable control coefficients.

A strategy to determine flux control coefficients from several protein levels has been developed by using adenovirus-mediated glucose-6-phosphatase (G6Pase) overexpression under the control of the cytomegalovirus promoter in rat hepatocytes. A 2-fold G6Pase overexpression did not alter $C^{\text{Glycolysis}}_{\text{G6Pase}}$ or $C^{\text{Glycolysis}}_{\text{GK}}$ (GK, glucokinase). However, if G6Pase is overexpressed by 4 folds, then $C^{\text{Glycogen-synthesis}}_{\text{GK}}$ diminished from 2.8 to 1.8 and there was a 35% lowering in glycogen synthesis [150]. However, this approach allows titration of flux only above the basal enzyme activities found in the cell, but not below.

These experimental inconveniences have been circumvented by using inducible gene expression systems based in the *lac*, Lambda, nisin, GAL, tetracycline, and other inducible promoters, in bacteria and yeast [151, 152]. However, a problem frequently encountered with inducible promoters is that a steady-state of protein expression is difficult to attain [151, 152].

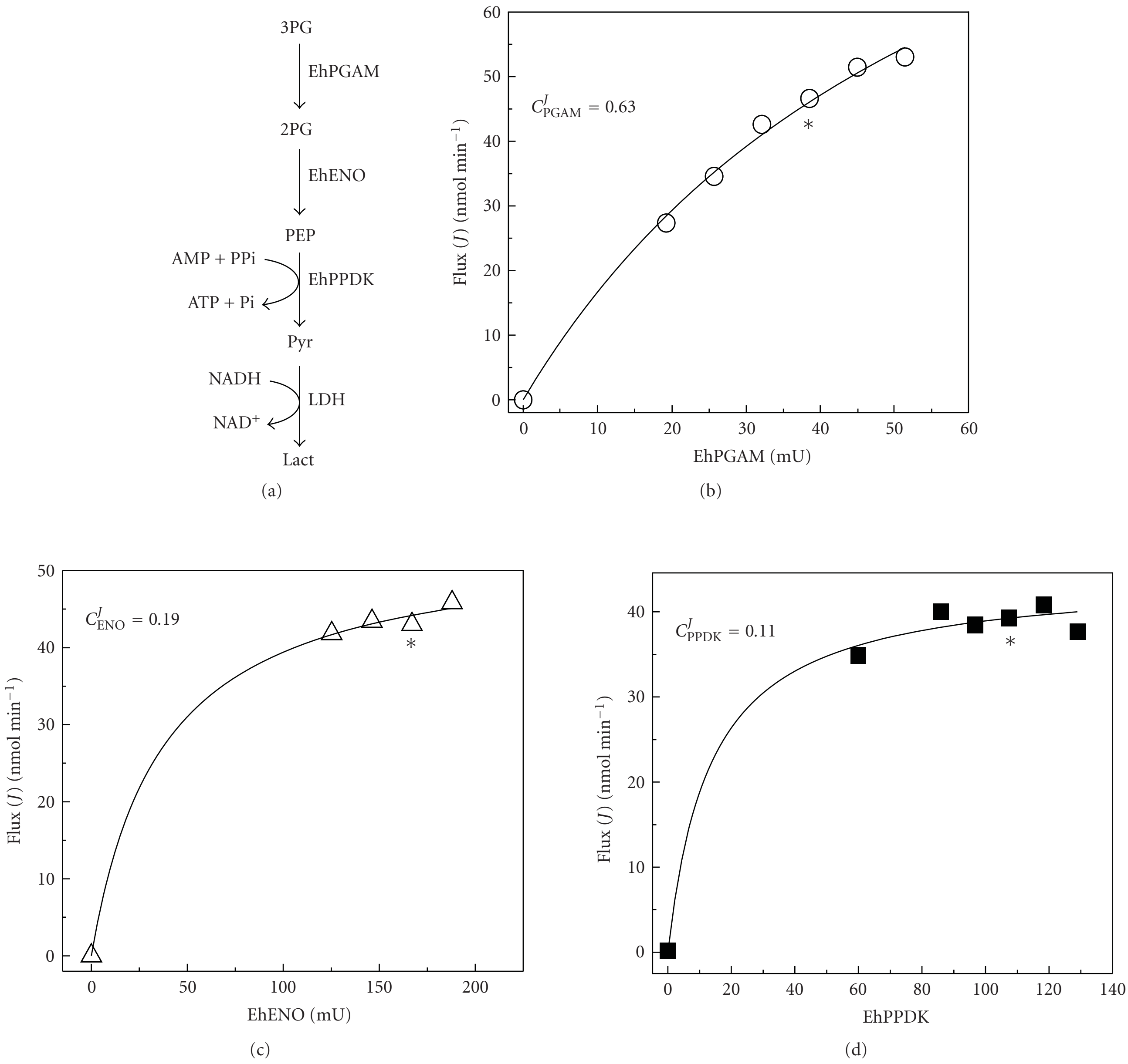


Figure 12: Determination of flux control coefficients in an in vitro reconstitution of the final section of *Entamoeba histolytica* glycolysis. Enzymatic assay with the three recombinant enzymes from the ameba: EhPGAM, EhENO, and EhPPDK. LDH, commercial lactate dehydrogenase. The flux control coefficient was determined at the *marked position. 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate. Modified from [143].

Recently, Jensen and Hammer described the design of synthetic promoter libraries (SPL), in particular for *L. lactis* metabolic optimization [153]. These promoters maintain constant the array of the known consensus sequences for *L. lactis* gene transcription (−10 and −35 boxes), while the nucleotide sequence between these boxes (a spacer sequence of 17 ± 1 bp) is randomized, thus producing a set of promoters with different transcriptional strength. These promoter libraries allow the transcription and protein expression several folds above and below the wild-type levels of enzyme activity [153], thus enhancing the usefulness of this approach for MCA studies.

The control distribution of glycolysis in *E. coli* and *L. lactis*, as discussed in Section 3.2 [17, 24, 27, 151], has been determined by using the SPL technology. SPL for yeast, mammalian and plant cells are also under development [151, 152]. Certainly, the advances in genetic engineering in combination with MCA allow better experimental designs for metabolic optimization of micro-organisms of biotechnological interest.

*Concluding remarks*

(1) The frequently recurred idea of manipulating the key enzyme or rate-limiting step (a concept based on a qualitative and rather intuitive background) to change metabolism is incorrect. As MCA has demonstrated, flux control is shared by multiple steps and it is not usually localized in only one step. MCA determines quantitatively the control that a given enzyme exerts on the flux and on intermediary concentration and helps to explain why an enzyme does or does not exert control.

(2) A metabolic pathway is manipulated to change the rate of the end-product formation (i.e., the flux) or the concentration of a relevant intermediary. As it is demonstrated in many unsuccessful experiments, it is not enough to overexpress one enzyme (the rate-limiting step) or many arbitrarily selected sites of the pathway. MCA proposes an initial experimental analysis that determines the structural control of the pathway and identifies the sites (enzymes and transporters) with higher control coefficients values (i.e., targets to be manipulated). For example, if there is a system composed of six enzymes and three of them have flux control coefficients with values of 0.2 or higher and the other three with values of 0.1 or lower, the three enzymes with high control coefficients must be overexpressed (if a flux increase is desired) or repressed (if flux inhibition is the objective) and not only one of them. If one of the selected enzymes is strongly inhibited by its product or has allosteric inhibition, the overexpression of this enzyme might not be enough to increase the flux, as it may also be necessary to moderately vary the product and allosteric modulator consuming enzymes.

(3) If the aim of the researcher is a metabolite concentration increase, which is not the end product of the pathway, MCA suggests the overexpression of those enzymes or transporters in the supply block with the highest control coefficients and/or the repression of those enzymes in the demand block with the highest control coefficients. These manipulations may become complicated if the metabolite of interest has allosteric interactions with enzymes and transporters (inhibition and activation) of both the supply and demand blocks. It is recalled that ethanol production in yeast and lactate and acetate production in lactobacteria do not increase by overexpressing PFK-1, an allosteric enzyme and the presumed rate-limiting step of glycolysis. In fact, the flux was diminished with an excessive PFK-1 overexpression. However, the analysis of these results reveals that the F1,6BP concentration is indeed increased many times over the control level. Another strategy for eliminating the feedback inhibition might be the introduction of mutations on the enzymes that are closer to the metabolite of interest.

## ABBREVIATIONS

| | |
|---|---|
| ADH: | alcohol dehydrogenase |
| CK: | creatine kinase |
| ENO: | enolase |
| GAPDH: | glyceraldehyde-3 phosphate dehydrogenase |
| HPI: | hexose phosphate isomerase |
| LDH: | lactate dehydrogenase |
| PDC: | pyruvate decarboxylase |
| PGK: | phosphoglucokinase |
| PGAM: | phosphoglycerate mutase |
| TPI: | triose phosphate isomerase |
| PPi-PFK: | pyrophosphate-dependent phosphofructokinase |
| $\alpha$-GPDH: | $\alpha$-glycerophosphate dehydrogenase |
| F6B: | fructose-6-phosphate |
| F1,6BP: | fructose-1,6-bisphosphate |
| G1P: | glucose-1-phosphate |
| G6P: | glucose-6-phosphate |
| GSH: | reduced glutathione |
| $\gamma$-EC: | $\gamma$-glutamylcysteine |
| MCA: | metabolic control analysis |
| siRNA: | small interfering RNA. |

## ACKNOWLEDGMENT

The present work was partially supported by CONACyT-Mexico Grants no. 60517 to RMS and 46719-Q to ES.

## REFERENCES

[1] E. A. Newsholme and C. Start, *Regulation of Metabolism*, John Wiley & Sons, London, UK, 1973.

[2] D. Fell, *Understanding the Control of Metabolism*, Portland Press, London, UK, 1997.

[3] M. David, I. R. Rasched, and H. Sund, "Studies of glutamate dehydrogenase. Methionine-169: the preferentially carboxymethylated residue," *European Journal of Biochemistry*, vol. 74, no. 2, pp. 379–385, 1977.

[4] H. Teng, E. Segura, and C. Grubmeyer, "Conserved cysteine residues of histidinol dehydrogenase are not involved in catalysis. Novel chemistry required for enzymatic aldehyde oxidation," *Journal of Biological Chemistry*, vol. 268, no. 19, pp. 14182–14188, 1993.

[5] T. N. C. Wells, M. A. Payton, and A. E. Proudfoot, "Inhibition of phosphomannose isomerase by mercury ions," *Biochemistry*, vol. 33, no. 24, pp. 7641–7646, 1994.

[6] B. Poolman, B. Bosman, J. Kiers, and W. N. Konings, "Control of glycolysis by glyceraldehyde-3-phosphate dehydrogenase in *Streptococcus cremoris* and *Streptococcus lactis*," *Journal of Bacteriology*, vol. 169, no. 12, pp. 5887–5890, 1987.

[7] R. Moreno-Sánchez and M. E. Torres-Márquez, "Control of oxidative phosphorylation in mitochondria, cells and tissues," *International Journal of Biochemistry*, vol. 23, no. 11, pp. 1163–1174, 1991.

[8] K. A. Webster, "Evolution of the coordinate regulation of glycolytic enzyme genes by hypoxia," *Journal of Experimental Biology*, vol. 206, no. 17, pp. 2911–2922, 2003.

[9] J. Heinisch, "Isolation and characterization of the two structural genes coding for phosphofructokinase in yeast,"

*Molecular and General Genetics*, vol. 202, no. 1, pp. 75–82, 1986.

[10] S. E. C. Davies and K. M. Brindle, "Effects of overexpression of phosphofructokinase on glycolysis in the yeast *Saccharomyces cerevisiae*," *Biochemistry*, vol. 31, no. 19, pp. 4729–4735, 1992.

[11] J. Hauf, F. K. Zimmermann, and S. Müller, "Simultaneous genomic overexpression of seven glycolytic enzymes in the yeast *Saccharomyces cerevisiae*," *Enzyme and Microbial Technology*, vol. 26, no. 9-10, pp. 688–698, 2000.

[12] I. Schaaff, J. Heinisch, and F. K. Zimmermann, "Overproduction of glycolytic enzymes in yeast," *Yeast*, vol. 5, no. 4, pp. 285–290, 1989.

[13] J. J. Hornberg, B. Binder, F. J. Bruggeman, B. Schoeberl, R. Heinrich, and H. V. Westerhoff, "Control of MAPK signalling: from complexity to what really matters," *Oncogene*, vol. 24, no. 36, pp. 5533–5542, 2005.

[14] H. P. Smits, J. Hauf, S. Müller, et al., "Simultaneous overexpression of enzymes of the lower part of glycolysis can enhance the fermentative capacity of *Saccharomyces cerevisiae*," *Yeast*, vol. 16, no. 14, pp. 1325–1334, 2000.

[15] M. Emmerling, J. E. Bailey, and U. Sauer, "Glucose catabolism of *Escherichia coli* strains with increased activity and altered regulation of key glycolytic enzymes," *Metabolic Engineering*, vol. 1, no. 2, pp. 117–127, 1999.

[16] M. Emmerling, J. E. Bailey, and U. Sauer, "Altered regulation of pyruvate kinase or co-overexpression of phosphofructokinase increases glycolytic fluxes in resting *Escherichia coli*," *Biotechnology and Bioengineering*, vol. 67, no. 5, pp. 623–627, 2000.

[17] C. Solem, B. J. Koebmann, and P. R. Jensen, "Glyceraldehyde-3-phosphate dehydrogenase has no control over glycolytic flux in *Lactococcus lactis* MG1363," *Journal of Bacteriology*, vol. 185, no. 5, pp. 1564–1571, 2003.

[18] G. J. G. Ruijter, H. Panneman, and J. Visser, "Overexpression of phosphofructokinase and pyruvate kinase in citric acid-producing *Aspergillus niger*," *Biochimica et Biophysica Acta*, vol. 1334, no. 2-3, pp. 317–326, 1997.

[19] A. M. Urbano, H. Gillham, Y. Groner, and K. M. Brindle, "Effects of overexpression of the liver subunit of 6-phosphofructo-1-kinase on the metabolism of a cultured mammalian cell line," *Biochemical Journal*, vol. 352, part 3, pp. 921–927, 2000.

[20] B. M. Bonini, P. Van Dijck, and J. M. Thevelein, "Uncoupling of the glucose growth defect and the deregulation of glycolysis in *Saccharomyces cerevisiae tps1* mutants expressing trehalose-6-phosphate-insensitive hexokinase from *Schizosaccharomyces pombe*," *Biochimica et Biophysica Acta*, vol. 1606, no. 1–3, pp. 83–93, 2003.

[21] J. Rivoal and A. D. Hanson, "Metabolic control of anaerobic glycolysis. Overexpression of lactate dehydrogenase in transgenic tomato roots supports the Davies-Roberts hypothesis and points to a critical role for lactate secretion," *Plant Physiology*, vol. 106, no. 3, pp. 1179–1185, 1994.

[22] S. Thomas, P. J. F. Mooney, M. M. Burrell, and D. A. Fell, "Metabolic control analysis of glycolysis in tuber tissue of potato (*Solanum tuberosum*): explanation for the low control coefficient of phosphofructokinase over respiratory flux," *Biochemical Journal*, vol. 322, part 1, pp. 119–127, 1997.

[23] R. M. O'Doherty, D. L. Lehman, J. Seoane, A.M. Gómez-Foix, J. J. Guinovart, and C. B. Newgard, "Differential metabolic effects of adenovirus-mediated glucokinase and hexokinase I overexpression in rat primary hepatocytes," *Journal of Biological Chemistry*, vol. 271, no. 34, pp. 20524–20530, 1996.

[24] H. W. Andersen, M. B. Pedersen, K. Hammer, and P. R. Jensen, "Lactate dehydrogenase has no control on lactate production but has a strong negative control on formate production in *Lactococcus lactis*," *European Journal of Biochemistry*, vol. 268, no. 24, pp. 6379–6389, 2001.

[25] B. J. Koebmann, C. Solem, and P. R. Jensen, "Control analysis as a tool to understand the formation of the *las* operon in *Lactococcus lactis*," *FEBS Journal*, vol. 272, no. 9, pp. 2292–2303, 2005.

[26] B. J. Koebmann, H. V. Westerhoff, J. L. Snoep, D. Nilsson, and P. R. Jensen, "The glycolytic flux in *Escherichia coli* is controlled by the demand for ATP," *Journal of Bacteriology*, vol. 184, no. 14, pp. 3909–3916, 2002.

[27] B. J. Koebmann, H. W. Andersen, C. Solem, and P. R. Jensen, "Experimental determination of control of glycolysis in *Lactococcus lactis*," *Antonie van Leeuwenhoek*, vol. 82, no. 1–4, pp. 237–248, 2002.

[28] B. J. Koebmann, C. Solem, M. B. Pedersen, D. Nilsson, and P. R. Jensen, "Expression of genes encoding $F_1$-ATPase results in uncoupling of glycolysis from biomass production in *Lactococcus lactis*," *Applied and Environmental Microbiology*, vol. 68, no. 9, pp. 4274–4282, 2002.

[29] S. Thomas and D. A. Fell, "A control analysis exploration of the role of ATP utilisation in glycolytic-flux control and glycolytic-metabolite-concentration regulation," *European Journal of Biochemistry*, vol. 258, no. 3, pp. 956–967, 1998.

[30] J.-H. S. Hofmeyr and A. Cornish-Bowden, "Regulating the cellular economy of supply and demand," *FEBS Letters*, vol. 476, no. 1-2, pp. 47–51, 2000.

[31] D. Mendoza-Cózatl, H. Loza-Tavera, A. Hernández-Navarro, and R. Moreno-Sánchez, "Sulfur assimilation and glutathione metabolism under cadmium stress in yeast, protists and plants," *FEMS Microbiology Reviews*, vol. 29, no. 4, pp. 653–671, 2005.

[32] A. Cornish-Bowden and M. L. Cárdenas, "Information transfer in metabolic pathways. Effects of irreversible steps in computer models," *European Journal of Biochemistry*, vol. 268, no. 24, pp. 6616–6624, 2001.

[33] A. Meister, "Glutathione metabolism," *Methods in Enzymology*, vol. 251, pp. 3–7, 1995.

[34] G. Noctor, A.-C. M. Arisi, L. Jouanin, and C. H. Foyer, "Manipulation of glutathione and amino acid biosynthesis in the chloroplast," *Plant Physiology*, vol. 118, no. 2, pp. 471–482, 1998.

[35] Y. L. Zhu, E. A. H. Pilon-Smits, A. S. Tarun, S. U. Weber, L. Jouanin, and N. Terry, "Cadmium tolerance and accumulation in Indian mustard is enhanced by overexpressing $\gamma$-glutamylcysteine synthetase," *Plant Physiology*, vol. 121, no. 4, pp. 1169–1177, 1999.

[36] S. Lee, J. S. Moon, T.-S. Ko, D. Petros, P. B. Goldsbrough, and S. S. Korban, "Overexpression of Arabidopsis phytochelatin synthase paradoxically leads to hypersensitivity to cadmium stress," *Plant Physiology*, vol. 131, no. 2, pp. 656–663, 2003.

[37] E. A. H. Pilon-Smits, S. Hwang, C. Mel Lytle, et al., "Overexpression of ATP sulfurylase in Indian mustard leads to increased selenate uptake, reduction, and tolerance," *Plant Physiology*, vol. 119, no. 1, pp. 123–132, 1999.

[38] Y. Hatzfeld, N. Cathala, C. Grignon, and J.-C. Davidian, "Effect of ATP sulfurylase overexpression in bright yellow 2 tobacco cells: regulation of ATP sulfurylase and $So_4^{2-}$ transport activities," *Plant Physiology*, vol. 116, no. 4, pp. 1307–1313, 1998.

[39] K. Saito, M. Kurosawa, K. Tatsuguchi, Y. Takagi, and I. Murakoshi, "Modulation of cysteine biosynthesis in chloroplasts of transgenic tobacco overexpressing cysteine synthase [*O*-acetylserine(thiol)-lyase]," *Plant Physiology*, vol. 106, no. 3, pp. 887–895, 1994.

[40] K. Harms, P. von Ballmoos, C. Brunold, R. Höfgen, and H. Hesse, "Expression of a bacterial serine acetyltransferase in transgenic potato plants leads to increased levels of cysteine and glutathione," *The Plant Journal*, vol. 22, no. 4, pp. 335–343, 2000.

[41] Y. Inoue, K.-I. Sugiyama, S. Izawa, and A. Kimura, "Molecular identification of glutathione synthetase (*GSH2*) gene from *Saccharomyces cerevisiae*," *Biochimica et Biophysica Acta*, vol. 1395, no. 3, pp. 315–320, 1998.

[42] S.-J. Kim, Y. H. Shin, K. Kim, E.-H. Park, J.-H. Sa, and C.-J. Lim, "Regulation of the gene encoding glutathione synthetase from the fission yeast," *Journal of Biochemistry and Molecular Biology*, vol. 36, no. 3, pp. 326–331, 2003.

[43] Y. L. Zhu, E. A. H. Pilon-Smits, L. Jouanin, and N. Terry, "Overexpression of glutathione synthetase in Indian mustard enhances cadmium accumulation and tolerance," *Plant Physiology*, vol. 119, no. 1, pp. 73–79, 1999.

[44] G. Creissen, J. Firmin, M. Fryer, et al., "Elevated glutathione biosynthetic capacity in the chloroplasts of transgenic tobacco plants paradoxically causes increased oxidative stress," *The Plant Cell*, vol. 11, no. 7, pp. 1277–1291, 1999.

[45] C. M. Grant, F. H. MacIver, and I. W. Dawes, "Glutathione synthetase is dispensable for growth under both normal and oxidative stress conditions in the yeast *Saccharomyces cerevisiae* due to an accumulation of the dipeptide $\gamma$-glutamylcysteine," *Molecular Biology of the Cell*, vol. 8, no. 9, pp. 1699–1707, 1997.

[46] D. F. Ortiz, T. Ruscitti, K. F. McCue, and D. W. Ow, "Transport of metal-binding peptides by HMT1, a fission yeast ABC-type vacuolar membrane protein," *Journal of Biological Chemistry*, vol. 270, no. 9, pp. 4721–4728, 1995.

[47] P. Niederberger, R. Prasad, G. Miozzari, and H. Kacser, "A strategy for increasing an in vivo flux by genetic manipulations: the tryptophan system of yeast," *Biochemical Journal*, vol. 287, part 2, pp. 473–479, 1992.

[48] R. Katsumata and M. Ikeda, "Hyperproduction of tryptophan in *Corynebacterium glutamicum* by pathway engineering," *Nature Biotechnology*, vol. 11, no. 8, pp. 921–925, 1993.

[49] S. Morbach, H. Sahm, and L. Eggeling, "Use of feedback-resistant threonine dehydratases of *Corynebacterium glutamicum* to increase carbon flux towards L-isoleucine," *Applied and Environmental Microbiology*, vol. 61, no. 12, pp. 4315–4320, 1995.

[50] M. A. G. Koffas, G. Y. Jung, J. C. Aon, and G. Stephanopoulos, "Effect of pyruvate carboxylase overexpression on the physiology of *Corynebacterium glutamicum*," *Applied and Environmental Microbiology*, vol. 68, no. 11, pp. 5422–5428, 2002.

[51] L. Padilla, R. Krämer, G. Stephanopoulos, and E. Agosin, "Overproduction of trehalose: heterologous expression of *Escherichia coli* trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase in *Corynebacterium glutamicum*," *Applied and Environmental Microbiology*, vol. 70, no. 1, pp. 370–376, 2004.

[52] E. Radmacher, A. Vaitsikova, U. Burger, K. Krumbach, H. Sahm, and L. Eggeling, "Linking central metabolism with increased pathway flux: L-valine accumulation by *Corynebacterium glutamicum*," *Applied and Environmental Microbiology*, vol. 68, no. 5, pp. 2246–2250, 2002.

[53] P. Simic, J. Willuhn, H. Sahm, and L. Eggeling, "Identification of *glyA* (encoding serine hydroxymethyltransferase) and its use together with the exporter ThrE to increase L-threonine accumulation by *Corynebacterium glutamicum*," *Applied and Environmental Microbiology*, vol. 68, no. 7, pp. 3321–3327, 2002.

[54] H. W. Wisselink, A. P. H. A. Moers, A. E. Mars, M. H. N. Hoefnagel, W. M. de Vos, and J. Hugenholtz, "Overproduction of heterologous mannitol 1-phosphatase: a key factor for engineering mannitol production by *Lactococcus lactis*," *Applied and Environmental Microbiology*, vol. 71, no. 3, pp. 1507–1514, 2005.

[55] V. Ladero, A. Ramos, A. Wiersma, et al., "High-level production of the low-calorie sugar sorbitol by *Lactobacillus plantarum* through metabolic engineering," *Applied and Environmental Microbiology*, vol. 73, no. 6, pp. 1864–1872, 2007.

[56] C. Solem, B. J. Koebmann, F. Yang, and P. R. Jensen, "The *las* enzymes control pyruvate metabolism in *Lactococcus lactis* during growth on maltose," *Journal of Bacteriology*, vol. 189, no. 18, pp. 6727–6730, 2007.

[57] R. Moreno-Sánchez, S. Rodríguez-Enríquez, A. Marín-Hernández, and E. Saavedra, "Energy metabolism in tumor cells," *FEBS Journal*, vol. 274, no. 6, pp. 1393–1418, 2007.

[58] B. Altenberg and K. O. Greulich, "Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes," *Genomics*, vol. 84, no. 6, pp. 1014–1020, 2004.

[59] D. J. Discher, N. H. Bishopric, X. Wu, C. A. Peterson, and K. A. Webster, "Hypoxia regulates $\beta$-enolase and pyruvate kinase-M promoters by modulating Sp1/Sp3 binding to a conserved GC element," *Journal of Biological Chemistry*, vol. 273, no. 40, pp. 26087–26093, 1998.

[60] M. Egea, I. Metón, and I. V. Baanante, "Sp1 and Sp3 regulate glucokinase gene transcription in the liver of gilthead sea bream (*Sparus aurata*)," *Journal of Molecular Endocrinology*, vol. 38, no. 3-4, pp. 481–492, 2007.

[61] B. J. Murphy, G. K. Andrews, D. Bittel, et al., "Activation of metallothionein gene expression by hypoxia involves metal response elements and metal transcription factor-1," *Cancer Research*, vol. 59, no. 6, pp. 1315–1322, 1999.

[62] S. R. Riddle, A. Ahmad, S. Ahmad, et al., "Hypoxia induces hexokinase II gene expression in human lung cell line A549," *American Journal of Physiology*, vol. 278, no. 2, pp. L407–L416, 2000.

[63] M. L. Parolin, L. L. Spriet, E. Hultman, M. G. Hollidge-Horvat, N. L. Jones, and G. J. F. Heigenhauser, "Regulation of glycogen phosphorylase and PDH during exercise in human skeletal muscle during hypoxia," *American Journal of Physiology*, vol. 278, no. 3, pp. E522–E534, 2000.

[64] P. A. M. Michels, F. Bringaud, M. Herman, and V. Hannaert, "Metabolic functions of glycosomes in trypanosomatids," *Biochimica et Biophysica Acta*, vol. 1763, no. 12, pp. 1463–1477, 2006.

[65] F. R. Opperdoes and P. A. M. Michels, "Enzymes of carbohydrate metabolism as potential drug targets," *International Journal for Parasitology*, vol. 31, no. 5-6, pp. 481–489, 2001.

[66] L. Azema, S. Claustre, I. Alric, et al., "Interaction of substituted hexose analogues with the *Trypanosoma brucei* hexose transporter," *Biochemical Pharmacology*, vol. 67, no. 3, pp. 459–467, 2004.

[67] F. Lakhdar-Ghazal, C. Blonski, M. Willson, P. Michels, and J. Perie, "Glycolysis and proteases as targets for the design of new anti-trypanosome drugs," *Current Topics in Medicinal Chemistry*, vol. 2, no. 5, pp. 439–456, 2002.

[68] B. M. Bakker, H. V. Westerhoff, F. R. Opperdoes, and P. A. M. Michels, "Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs," *Molecular and Biochemical Parasitology*, vol. 106, no. 1, pp. 1–10, 2000.

[69] A. H. Fairlamb and A. Cerami, "Metabolism and functions of trypanothione in the Kinetoplastida," *Annual Review of Microbiology*, vol. 46, pp. 695–729, 1992.

[70] S. Müller, E. Liebau, R. D. Walter, and R. L. Krauth-Siegel, "Thiol-based redox metabolism of protozoan parasites," *Trends in Parasitology*, vol. 19, no. 7, pp. 320–328, 2003.

[71] S. A. Le Quesne and A. H. Fairlamb, "Regulation of a high-affinity diamine transport system in *Trypanosoma cruzi* epimastigotes," *Biochemical Journal*, vol. 316, part 2, pp. 481–486, 1996.

[72] C. Dumas, M. Ouellette, J. Tovar, et al., "Disruption of the trypanothione reductase gene of *Leishmania* decreases its ability to survive oxidative stress in macrophages," *The EMBO Journal*, vol. 16, no. 10, pp. 2590–2598, 1997.

[73] J. Tovar, M. L. Cunningham, A. C. Smith, S. L. Croft, and A. H. Fairlamb, "Down-regulation of *Leishmania donovani* trypanothione reductase by heterologous expression of a trans-dominant mutant homologue: effect on parasite intracellular survival," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 5311–5316, 1998.

[74] J. Tovar, S. Wilkinson, J. C. Mottram, and A. H. Fairlamb, "Evidence that trypanothione reductase is an essential enzyme in *Leishmania* by targeted replacement of the *tryA* gene locus," *Molecular Microbiology*, vol. 29, no. 2, pp. 653–660, 1998.

[75] S. Krieger, W. Schwarz, M. R. Arlyanayagam, A. H. Fairlamb, R. L. Krauth-Siegel, and C. Clayton, "Trypanosomes lacking trypanothione reductase are avirulent and show increased sensitivity to oxidative stress," *Molecular Microbiology*, vol. 35, no. 3, pp. 542–552, 2000.

[76] J. M. Kelly, M. C. Taylor, K. Smith, K. J. Hunter, and A. H. Fairlamb, "Phenotype of recombinant *Leishmania donovani* and *Trypanosoma cruzi* which over-express trypanothione reductase. Sensitivity towards agents that are thought to induce oxidative stress," *European Journal of Biochemistry*, vol. 218, no. 1, pp. 29–37, 1993.

[77] M. A. Comini, S. A. Guerrero, S. Haile, U. Menge, H. Lünsdorf, and L. Flohé, "Validation of *Trypanosoma brucei* trypanothione synthetase as drug target," *Free Radical Biology and Medicine*, vol. 36, no. 10, pp. 1289–1302, 2004.

[78] M. R. Ariyanayagam, S. L. Oza, M. L. Guther, and A. H. Fairlamb, "Phenotypic analysis of trypanothione synthetase knockdown in the African trypanosome," *Biochemical Journal*, vol. 391, part 2, pp. 425–432, 2005.

[79] T. T. Huynh, V. T. Huynh, M. A. Harmon, and M. A. Phillips, "Gene knockdown of $\gamma$-glutamylcysteine synthetase by $\text{RNA}_\text{i}$ in the parasitic protozoa *Trypanosoma brucei* demonstrates that it is an essential enzyme," *Journal of Biological Chemistry*, vol. 278, no. 41, pp. 39794–39800, 2003.

[80] C. Guimond, N. Trudel, C. Brochu, et al., "Modulation of gene expression in *Leishmania* drug resistant mutants as determined by targeted DNA microarrays," *Nucleic Acids Research*, vol. 31, no. 20, pp. 5886–5896, 2003.

[81] S. K. Shahi, R. L. Krauth-Siegel, and C. E. Clayton, "Overexpression of the putative thiol conjugate transporter *Tb*MRPA causes melarsoprol resistance in *Trypanosoma brucei*," *Molecular Microbiology*, vol. 43, no. 5, pp. 1129–1138, 2002.

[82] H. Kacser and J. A. Burns, "The control of flux," *Symposia of the Society for Experimental Biology*, vol. 27, pp. 65–104, 1973.

[83] H. Kacser and J. A. Burns, "Molecular democracy: who shares the controls?" *Biochemical Society Transactions*, vol. 7, no. 5, pp. 1149–1160, 1979.

[84] R. Heinrich, S. M. Rapoport, and T. A. Rapoport, "Metabolic regulation and mathematical models," *Progress in Biophysics and Molecular Biology*, vol. 32, pp. 1–82, 1978.

[85] T. A. Rapoport, R. Heinrich, G. Jacobasch, and S. Rapoport, "A linear steady state treatment of enzymatic chains. A mathematical model of glycolysis of human erythrocytes," *European Journal of Biochemistry*, vol. 42, no. 1, pp. 107–120, 1974.

[86] H. J. Flint, R. W. Tateson, I. B. Barthelmess, D. J. Porteous, W. D. Donachie, and H. Kacser, "Control of the flux in the arginine pathway of *Neurospora crassa*. Modulations of enzyme activity and concentration," *Biochemical Journal*, vol. 200, no. 2, pp. 231–246, 1981.

[87] N. S. Cohen, C.-W. Cheung, E. Sijuwade, and L. Raijman, "Kinetic properties of carbamoyl-phosphate synthase (ammonia) and ornithine carbamoyltransferase in permeabilized mitochondria," *Biochemical Journal*, vol. 282, part 1, pp. 173–180, 1992.

[88] S. G. Powers-Lee, R. A. Mastico, and M. Bendayan, "The interaction of rat liver carbamoyl phosphate synthetase and ornithine transcarbamoylase with inner mitochondrial membranes," *Journal of Biological Chemistry*, vol. 262, no. 32, pp. 15683–15688, 1987.

[89] R. J. Middleton and H. Kacser, "Enzyme variation, metabolic flux and fitness: alcohol dehydrogenase in *Drosophila melanogaster*," *Genetics*, vol. 105, no. 3, pp. 633–650, 1983.

[90] S. Rodríguez-Enríquez, M. E. Torres-Márquez, and R. Moreno-Sánchez, "Substrate oxidation and ATP sypply in AS-30D hepatoma cells," *Archives of Biochemistry and Biophysics*, vol. 375, no. 1, pp. 21–30, 2000.

[91] C. Garcia, J. P. Pardo, and R. Moreno-Sánchez, "Control of oxidative phosphorylation supported by NAD-linked substrates in rat brain mitochondria," *Biochemical Archives*, vol. 12, no. 3, pp. 157–176, 1996.

[92] F. N. Gellerich, W. S. Kunz, and R. Bohnensack, "Estimation of flux control coefficients from inhibitor titrations by non-linear regression," *FEBS Letters*, vol. 274, no. 1-2, pp. 167–170, 1990.

[93] R. Moreno-Sánchez, S. Devars, F. López-Gómez, A. Uribe, and N. Corona, "Distribution of control of oxidative phosphorylation in mitochondria oxidizing NAD-linked substrates," *Biochimica et Biophysica Acta*, vol. 1060, no. 3, pp. 284–292, 1991.

[94] R. Moreno-Sánchez, "Contribution of the translocator of adenine nucleotides and the ATP synthase to the control of oxidative phosphorylation and arsenylation in liver mitochondria," *Journal of Biological Chemistry*, vol. 260, no. 23, pp. 12554–12560, 1985.

[95] R. Rossignol, T. Letellier, M. Malgat, C. Rocher, and J.-P. Mazat, "Tissue variation in the control of oxidative phosphorylation: implication for mitochondrial diseases," *Biochemical Journal*, vol. 347, part 1, pp. 45–53, 2000.

[96] F. J. López-Gómez, M. E. Torres-Márquez, and R. Moreno-Sánchez, "Control of oxidative phosphorylation in AS-30D hepatoma mitochondria," *International Journal of Biochemistry*, vol. 25, no. 3, pp. 373–377, 1993.

[97] E. Wisniewski, W. S. Kunz, and F. N. Gellerich, "Phosphate affects the distribution of flux control among the enzymes of oxidative phosphorylation in rat skeletal muscle

mitochondria," *Journal of Biological Chemistry*, vol. 268, no. 13, pp. 9343–9346, 1993.

[98] T. Latsis, B. Andersen, and L. Agius, "Diverse effects of two allosteric inhibitors on the phosphorylation state of glycogen phosphorylase in hepatocytes," *Biochemical Journal*, vol. 368, part 1, pp. 309–316, 2002.

[99] S. A. Gupte, M. Arshad, S. Viola, et al., "Pentose phosphate pathway coordinates multiple redox-controlled relaxing mechanisms in bovine coronary arteries," *American Journal of Physiology*, vol. 285, no. 6, pp. H2316–H2326, 2003.

[100] D. C. Wallace, "Mitochondrial diseases in man and mouse," *Science*, vol. 283, no. 5407, pp. 1482–1488, 1999.

[101] X. L. Zu and M. Guppy, "Cancer metabolismml: facts, fantasy, and fiction," *Biochemical and Biophysical Research Communications*, vol. 313, no. 3, pp. 459–465, 2004.

[102] M. Erecinska and F. Dagani, "Relationships between the neuronal sodium/potassium pump and energy metabolism. Effects of $K^+$, $Na^+$, and adenosine triphosphate in isolated brain synaptosomes," *Journal of General Physiology*, vol. 95, no. 4, pp. 591–616, 1990.

[103] G. C. Brown, P. L. Lakin-Thomas, and M. D. Brand, "Control of respiration and oxidative phosphorylation in isolated rat liver cells," *European Journal of Biochemistry*, vol. 192, no. 2, pp. 355–362, 1990.

[104] R. P. Hafner, G. C. Brown, and M. D. Brand, "Analysis of the control of respiration rate, phosphorylation rate, proton leak rate and protonmotive force in isolated mitochondria using the 'top-down' approach of metabolic control theory," *European Journal of Biochemistry*, vol. 188, no. 2, pp. 313–319, 1990.

[105] R. J. Wanders, A. K. Groen, C. W. van Roermund, and J. M. Tager, "Factors determining the relative contribution of the adenine-nucleotide translocator and the ADP-regenerating system to the control of oxidative phosphorylation in isolated rat-liver mitochondria," *European Journal of Biochemistry*, vol. 142, no. 2, pp. 417–424, 1984.

[106] A. K. Groen, C. W. T. van Roermund, R. C. Vervoorn, and J. M. Tager, "Control of gluconeogenesis in rat liver cells. Flux control coefficients of the enzymes in the gluconeogenic pathway in the absence and presence of glucagon," *Biochemical Journal*, vol. 237, part 2, pp. 379–389, 1986.

[107] A. Marín-Hernández, S. Rodríguez-Enríquez, P. A. Vital-González, et al., "Determining and understanding the control of glycolysis in fast-growth tumor cells: flux control by an over-expressed but strongly product-inhibited hexokinase," *FEBS Journal*, vol. 273, no. 9, pp. 1975–1988, 2006.

[108] A. L. Kruckeberg, H. E. Neuhaus, R. Feil, L. D. Gottlieb, and M. Stitt, "Decreased-activity mutants of phosphoglucose isomerase in the cytosol and chloroplast of *Clarkia xantiana*. Impact on mass-action ratios and fluxes to sucrose and starch, and estimation of flux control coefficients and elasticity coefficients," *Biochemical Journal*, vol. 261, part 2, pp. 457–467, 1989.

[109] P. A. Quant, D. Robin, P. Robin, J. Girard, and M. D. Brand, "A top-down control analysis in isolated rat liver mitochondria: can the 3-hydroxy-3-methylglutaryl-CoA pathway be rate-controlling for ketogenesis?" *Biochimica et Biophysica Acta*, vol. 1156, no. 2, pp. 135–143, 1993.

[110] D. A. Fell and K. Snell, "Control analysis of mammalian serine biosynthesis. Feedback inhibition of the final step," *Biochemical Journal*, vol. 256, part 1, pp. 97–101, 1988.

[111] C. Chassagnole, D. A. Fell, B. Rais, B. Kudla, and J.-P. Mazat, "Control of the threonine-synthesis pathway in *Escherichia coli*: a theoretical and experimental approach," *Biochemical Journal*, vol. 356, part 2, pp. 433–444, 2001.

[112] J. L. Galazzo and J. E. Bailey, "Fermentation pathway kinetics and metabolic flux control in suspended and immobilized *Saccharomyces cerevisiae*," *Enzyme and Microbial Technology*, vol. 12, no. 3, pp. 162–172, 1990.

[113] J. A. Diderich, B. Teusink, J. Valkier, et al., "Strategies to determine the extent of control exerted by glucose transport on glycolytic flux in the yeast *Saccharomyces bayanus*," *Microbiology*, vol. 145, no. 12, pp. 3447–3454, 1999.

[114] H. V. Westerhoff and D. Kahn, "Control involving metabolism and gene expression: the square-matrix method for modular decomposition," *Acta Biotheoretica*, vol. 41, no. 1-2, pp. 75–83, 1993.

[115] J. R. Chase, D. L. Rothman, and R. G. Shulman, "Flux control in the rat gastrocnemius glycogen synthesis pathway by in vivo 13C/31P NMR spectroscopy," *American Journal of Physiology*, vol. 280, no. 4, pp. E598–E607, 2001.

[116] P. De Atauri, D. Orrell, S. Ramsey, and H. Bolouri, "Is the regulation of galactose 1-phosphate tuned against gene expression noise?" *Biochemical Journal*, vol. 387, part 1, pp. 77–84, 2005.

[117] M. G. Poolman, H. E. Assmus, and D. A. Fell, "Applications of metabolic modelling to plant metabolism," *Journal of Experimental Botany*, vol. 55, no. 400, pp. 1177–1186, 2004.

[118] P. Mendes, "GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems," *Computer Applications in the Biosciences*, vol. 9, no. 5, pp. 563–571, 1993.

[119] A. Cornish-Bowden and J.-H. S. Hofmeyr, "MetaModel: a program for modelling and control analysis of metabolic pathways on the IBM PC and compatibles," *Computer Applications in the Biosciences*, vol. 7, no. 1, pp. 89–93, 1991.

[120] H. M. Sauro, "SCAMP: a general-purpose simulator and metabolic control analysis program," *Computer Applications in the Biosciences*, vol. 9, no. 4, pp. 441–450, 1993.

[121] H. M. Sauro, "JARNAC: a system for interactive metabolic analysis," in *Animating the Cellular Map*, J.-H. S. Hofmeyr, J. M. Rohwer, and J. L. Snoep, Eds., pp. 221–228, Stellenbosch University Press, Stellenbosch, South Africa, 2000.

[122] B. G. Olivier, J. M. Rohwer, and J.-H. S. Hofmeyr, "Modelling cellular systems with PySCeS," *Bioinformatics*, vol. 21, no. 4, pp. 560–561, 2005.

[123] B. Teusink, J. Passarge, C. A. Reijenga, et al., "Can yeast glycolysis be understood terms of vitro kinetics of the constituent enzymes? Testing biochemistry," *European Journal of Biochemistry*, vol. 267, no. 17, pp. 5313–5329, 2000.

[124] L. Pritchard and D. B. Kell, "Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis," *European Journal of Biochemistry*, vol. 269, no. 16, pp. 3894–3904, 2002.

[125] B. M. Bakker, P. A. M. Michels, F. R. Opperdoes, and H. V. Westerhoff, "Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes," *Journal of Biological Chemistry*, vol. 272, no. 6, pp. 3207–3215, 1997.

[126] B. M. Bakker, P. A. M. Michels, F. R. Opperdoes, and H. V. Westerhoff, "What controls glycolysis in bloodstream form *Trypanosoma brucei*?" *Journal of Biological Chemistry*, vol. 274, no. 21, pp. 14551–14559, 1999.

[127] K. R. Albe and B. E. Wright, "Carbohydrate metabolism in *dictyostelium discoideumml*: II. Systems' analysis," *Journal of Theoretical Biology*, vol. 169, no. 3, pp. 243–251, 1994.

[128] J. M. Rohwer and F. C. Botha, "Analysis of sucrose accumulation in the sugar cane culm on the basis of in vitro

kinetic data," *Biochemical Journal*, vol. 358, part 2, pp. 437–445, 2001.

[129] G. R. Cronwright, J. M. Rohwer, and B. A. Prior, "Metabolic control analysis of glycerol synthesis in *Saccharomyces cerevisiae*," *Applied and Environmental Microbiology*, vol. 68, no. 9, pp. 4448–4456, 2002.

[130] J. Nielsen and H. S. Jorgensen, "Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of *Penicillium chrysogenum*," *Biotechnology Progress*, vol. 11, no. 3, pp. 299–305, 1995.

[131] M. G. Poolman, D. A. Fell, and S. Thomas, "Modelling photosynthesis and its control," *Journal of Experimental Botany*, vol. 51, pp. 319–328, 2000.

[132] Q. Hua, C. Yang, and K. Shimizu, "Metabolic control analysis for lysine synthesis using *Corynebacterium glutamicum* and experimental verification," *Journal of Bioscience and Bioengineering*, vol. 90, no. 2, pp. 184–192, 2000.

[133] H. A. Berthon, P. W. Kuchel, and P. F. Nixon, "High control coefficient of transketolase in the nonoxidative pentose phosphate pathway of human erythrocytes: NMR, antibody, and computer simulation studies," *Biochemistry*, vol. 31, no. 51, pp. 12792–12798, 1992.

[134] M. J. L. de Groot, W. Prathumpai, J. Visser, and G. J. G. Ruijter, "Metabolic control analysis of *Aspergillus niger* L-arabinose catabolism," *Biotechnology Progress*, vol. 21, no. 6, pp. 1610–1616, 2005.

[135] M. H. N. Hoefnagel, M. J. C. Starrenburg, D. E. Martens, et al., "Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis," *Microbiology*, vol. 148, no. 4, pp. 1003–1013, 2002.

[136] D. G. Mendoza-Cózatl and R. Moreno-Sánchez, "Control of glutathione and phytochelatin synthesis under cadmium stress. Pathway modeling for plants," *Journal of Theoretical Biology*, vol. 238, no. 4, pp. 919–936, 2006.

[137] A. Meister, "Glutathione synthesis," in *The Enzymes*, P. D. Boyer, Ed., vol. 10, pp. 671–697, Academic Press, New York, NY, USA, 1994.

[138] D. M. Speiser, S. L. Abrahamson, G. Banuelos, and D. W. Ow, "*Brassica juncea* produces a phytochelatin-cadmium-sulfide complex," *Plant Physiology*, vol. 99, no. 3, pp. 817–821, 1992.

[139] W. J. G. Vande and D. W. Ow, "Accumulation of metal-binding peptides in fission yeast requires *hmt*2$^+$," *Molecular Microbiology*, vol. 42, no. 1, pp. 29–36, 2001.

[140] J. R. Small and H. Kacser, "Responses of metabolic systems to large changes in enzyme activities and effectors. 1. The linear treatment of unbranched chains," *European Journal of Biochemistry*, vol. 213, no. 1, pp. 613–624, 1993.

[141] R. E. Reeves, "Metabolism of *Entamoeba histolytica* Schaudinn, 1903," *Advances in Parasitology*, vol. 23, pp. 105–142, 1984.

[142] E. Saavedra, A. Marín-Hernández, R. Encalada, A. Olivos, G. Mendoza-Hernández, and R. Moreno-Sánchez, "Kinetic modeling can describe in vivo glycolysis in *Entamoeba histolytica*," *FEBS Journal*, vol. 274, no. 18, pp. 4922–4940, 2007.

[143] E. Saavedra, R. Encalada, E. Pineda, R. Jasso-Chávez, and R. Moreno-Sánchez, "Glycolysis in *Entamoeba histolytica*: biochemical characterization of recombinant glycolytic enzymes and flux control analysis," *FEBS Journal*, vol. 272, no. 7, pp. 1767–1783, 2005.

[144] N. V. Torres, R. Souto, and E. Meléndez-Hevia, "Study of the flux and transition time control coefficient profiles in a metabolic system in vitro and the effect of an external stimulator," *Biochemical Journal*, vol. 260, part 3, pp. 763–769, 1989.

[145] C. Giersch, "Determining elasticities from multiple measurements of flux rates and metabolite concentrations. Application of the multiple modulation mehod to a reconstituted pathway," *European Journal of Biochemistry*, vol. 227, no. 1-2, pp. 194–201, 1995.

[146] R. Moreno-Sánchez, R. Encalada, A. Marín-Hernández, and E. Saavedra, "Experimental validation of metabolic pathway modeling. An illustration with glycolytic segments from *Entamoeba histolytica*," *FEBS Journal*, vol. 275, no. 13, pp. 3454–3469, 2008.

[147] J. Kurreck, "Antisense technologies: improvement through novel chemical modifications," *European Journal of Biochemistry*, vol. 270, no. 8, pp. 1628–1644, 2003.

[148] M. Stitt, W. P. Quick, U. Schurr, E.-D. Schulze, S. R. Rodermel, and L. Bogorad, "Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with 'antisense' *rbcS* II. Flux-control coefficients for photosynthesis in varying light, $Co_2$, and air humidity," *Planta*, vol. 183, no. 4, pp. 555–566, 1991.

[149] M.-A. Albert, J. R. Haanstra, V. Hannaert, et al., "Experimental and *in silico* analyses of glycolytic flux control in bloodstream form *Trypanosoma brucei*," *Journal of Biological Chemistry*, vol. 280, no. 31, pp. 28306–28315, 2005.

[150] S. Aiston, K. Y. Trinh, A. J. Lange, C. B. Newgard, and L. Agius, "Glucose-6-phosphatase overexpression lowers glucose 6-phosphate and inhibits glycogen synthesis and glycolysis in hepatocytes without affecting glucokinase translocation," *Journal of Biological Chemistry*, vol. 274, no. 35, pp. 24559–24566, 1999.

[151] B. J. Koebmann, J. Tornøe, B. Johansson, and P. R. Jensen, "Experimental modulation of gene expression," in *Metabolic Engineering in the Post Genomic Era*, B. N. Kholodenko and H. V. Westerhoff, Eds., pp. 155–179, Horizon Bioscience, Norfolk, UK, 2004.

[152] K. Hammer, I. Mijakovic, and P. R. Jensen, "Synthetic promoter libraries—tuning of gene expression," *Trends in Biotechnology*, vol. 24, no. 2, pp. 53–55, 2006.

[153] P. R. Jensen and K. Hammer, "The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters," *Applied and Environmental Microbiology*, vol. 64, no. 1, pp. 82–87, 1998.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 742504, 10 pages
doi:10.1155/2008/742504

*Methodology Report*

# Formal Implementation of a Performance Evaluation Model for the Face Recognition System

**Yong-Nyuo Shin,[1] Jason Kim,[1] Yong-Jun Lee,[2] Woochang Shin,[3] and Jin-Young Choi[4]**

[1] *Korea Information Security Agency, 78 Karak-dong, Songpa-gu, Seoul 138-160, South Korea*
[2] *LG CNS, Hoehyun-dong, 2-ga, Jung-gu, Seoul 100-630, South Korea*
[3] *Department of Internet Information, College of Natural Science & Engineering, Seokyeong University,
16-1 Jungneung-dong, Sungbuk-ku, Seoul 136-704, South Korea*
[4] *Division of Computer Science and Engineering, College of Information and Communications, Korea University,
Anam-dong, Sungbuk-ku, Seoul 136-701, South Korea*

Correspondence should be addressed to Woochang Shin, wcshin@imail.skuniv.ac.kr

Received 5 October 2007; Accepted 25 October 2007

Recommended by Daniel Howard

Due to usability features, practical applications, and its lack of intrusiveness, face recognition technology, based on information, derived from individuals' facial features, has been attracting considerable attention recently. Reported recognition rates of commercialized face recognition systems cannot be admitted as official recognition rates, as they are based on assumptions that are beneficial to the specific system and face database. Therefore, performance evaluation methods and tools are necessary to objectively measure the accuracy and performance of any face recognition system. In this paper, we propose and formalize a performance evaluation model for the biometric recognition system, implementing an evaluation tool for face recognition systems based on the proposed model. Furthermore, we performed evaluations objectively by providing guidelines for the design and implementation of a performance evaluation system, formalizing the performance test process.

## 1. INTRODUCTION

Face recognition systems provide the benefit of collecting a large amount of biometric information in a relatively easy and cost-effective manner, because they do not require subjects to bring any part of their body in contact with the recognition device intentionally, which results in fewer repercussions and less inconvenience when collecting the biometric information. An additional advantage exists, that is, the widely deployed image acquisition equipment can be used without modification. In particular, various face recognition algorithms and commercial systems have been developed and proposed, and the marketability of face recognition systems has increased, with many immigration-related facilities such as air and sea ports in many countries anxious to introduce face recognition systems after the 9.11 terror attacks in the US. These benefits, and the perceived necessity of increased security, have led to a rising social demand for face recognition systems; and certified performance evaluation has become important as a means of evaluating these face recognition systems.

This paper proposes a performance evaluation model (PEM) to evaluate the performance of biometric recognition systems; and designs and implements a performance evaluation tool that enables comparison and evaluation of face recognition systems, based on the proposed PEM. The PEM is designed to be compatible with related international standards, and contributes to the consistency and enhanced reliability of the performance evaluation tool that is developed with reference to the model.

Section 2 outlines existing studies related to performance evaluation of face recognition systems. Section 3 proposes a PEM to evaluate the performance of face recognition systems. Section 4 describes the design and implementation of the performance evaluation tool, based on the proposed PEM. Section 5 compares the performance evaluation method, utilizing the performance evaluation tool proposed in this paper, with existing performance evaluation programs. The last section provides a conclusion, and suggests some remarks for future research.

## 2. RELATED STUDIES

The representative certified performance evaluation programs for facial recognition systems are FacE Recognition Technology (FERET) and Face Recognition Vendor Testing (FRVT). As has been used since 1993, FERET includes not only performance evaluation of facial recognition systems but also the development of algorithms and the collection of a face recognition database. Headed by the U.S. Department of Defense (USDoD), FERET is an evaluation tool that has been systematically executed from 1993 through 1997 by testing changes in the environment (e.g., size, posture, background, etc.), differences in the time when pictures are taken, and the performance of algorithms in processing a mass-volume database. In particular, the FERET performance evaluations have been general evaluations designed to measure algorithms at the level of research centers. The major purpose of the FERET performance evaluation tool has been to implement adaptations to the latest facial recognition technologies and their flexibility. Therefore, the FERET test is neither used to clearly measure the influence of algorithms on the performance of individual components nor to assess performance in fully organized scientific manners under all operating conditions of a system [1–3].

Face Recognition Vendor Testing (FRVT) was a performance test for the face recognition system that was implemented using three Face Recognition Technology (FERET) performance evaluations (1994, 1995, and 1996). The FERET program introduced the evaluation technique in the face recognition area, and developed the face recognition area at the earliest level (system prototype development). However, as face recognition technology matured from the prototype level to the commercial system level, FRVT 2000 measured the performance of these commercial systems, and evaluated how far the technology had evolved through comparison with the last FERET evaluation. The public began to pay more attention to face recognition technology in 2002. As a consequence, FRVT 2002 measured the degree of technical development since 2000, evaluated the large-size databases that were in use, and introduced new experiments to better understand the performance of face recognition. Size, difficulty, and complexity of this performance evaluation were on the rise as the evaluation theory as well as the face recognition technology grew. For example, FERET SEP96 performed just 14.5 million comparisons over a period of 72 hours, while FRVT 2000 carried out 192 million comparisons in 72 hours. In contrast, FRVT 2002 introduced an evaluation that made 15 billion comparisons in 264 hours [4–6].

Certified performance evaluation programs like FERET and FRVT were designed to measure the algorithm accuracy of face recognition systems. For these projects, a common face image database was provided for the test, face recognition was performed for a certain period of time according to the respective method, and the results were evaluated. However, this method provides evaluation only for face recognition technology vendors that participated in the program during the evaluation period. In particular, database items were limited to image size, target posture, image acquisition environment, and time, which left the problem that various conditions of algorithm evaluation were not satisfied dynamically. Moreover, the algorithm evaluation environment was commissioned to each of the face recognition system developers, creating the problem of inconsistency in establishing a performance evaluation system environment. Furthermore, additional tasks were required in order to determine the accuracy of each algorithm, and to analyze the algorithm implementation result again. Therefore, it is necessary to design an algorithm evaluation method that can resolve these problems, to build a standardized evaluation environment, and to automatically figure out the evaluation result of the algorithm whose performance is measured in this environment.

Table 1: Classification of factors affecting the performance of biometric system.

| Classification | Particular factors |
|---|---|
| Population demographics | Age, ethnic origin, gender, occupation, etc. |
| Application | Template aging, shooting time, User friendliness, user motivation, etc. |
| User physiology | Hair, baldness, illnesses and diseases, eyelashes, nail length, skin tone, etc. |
| User behavior | Facial expression, dialect, motion, posture, angle, distance, stress, nervousness, etc. |
| Environmental influences | Background features (color, shadow, sound, etc.), lighting strength and direction, weather, (temperature, humidity, rain, snow, etc.) |
| User appearance | Outfit (hat, sleeves, pants, dress shoes, etc.), band (bandage), contact lenses, makeup, glasses, fake nails, hairstyle, rings, tattoos, etc. |
| Sensor and hardware | Dirt on camera lenses, focus, sensor quality, sensor change, transmission channel, etc. |
| User interface | Feedback, instructions, supervising director |

### 2.1. Factors affecting performance evaluation

Results from the performance evaluation of facial recognition systems change in accordance with varying factors, such as lighting, posture, facial expression, and elapsed time. The JTC 1/SC 37/WG 5 International Standard [7] classifies those factors affecting the performance of biometric systems.

As outlined in Table 1, there are a number of factors affecting the performance of facial recognition systems, and such factors must be prudently taken into consideration when the probe and gallery are selected during the performance evaluation of these systems. Algorithm performance evaluation of biometric recognition technology is conducted in such a way that the standard gallery is trained or registered, and is then compared with the test biometric information (probe) to be recognized, after which the similarities between the two sets of information are measured.

Table 2: Classification of the KISA's database.

| Criterion | Content |
|---|---|
| Gender | Male or female |
| Age | Age |
| Birthplace | Gyeonggi-do, Seoul |
| Lighting color | Fluorescent lamp, Incandescent lamp |
| Lighting direction | Front, right/left 45°, right/left 90°, right/left 135°, 180° |
| Glasses direction | Front, Right/Left 45°, Right/Left 90° |
| Facial expression | No Expression, Smiling, Angry, Closed Eyes, Surprised |
| Posture | Front, Right/Left 15°, Right/Left 30°,<br>Right/Left 45° |
| Hair condition | Natural Hairstyle, Hair Band/s, Glasses |

Generally, basic factors such as posture, angle, facial expression, lighting brightness, and gender are considered in the construction of a face image database. However, records of the face image database need to be further subdivided in order to process the test under conditions similar to those prevalent in the real world.

The research facial database that was developed by Korea Information Security Agency (KISA) from 2002 to 2004 was used for performance evaluation [8]. Table 2 shows a classification of KISA's database.

## 3. PERFORMANCE EVALUATION MODEL (PEM)

Many factors must be considered when building a fair performance evaluation system for biometric recognition systems. For example, to evaluate the performance of face recognition systems, a database of facial information for use in face recognition should be collected, and performance evaluation items (changes in facial expression, lighting, etc.) as well as performance evaluation measurement criteria such as false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER) should be selected. The face recognition system to be evaluated and a standardized interface for the performance evaluation system should be designed and international standards need to be applied at each stage of performance evaluation, in order to enhance fairness and reliability.

Thus, the performance evaluation model (PEM) is created to analyze and arrange the criteria to be considered in building up the performance evaluation system, and to support the development of the performance evaluation system. The PEM presents the basic system structure, guidelines, and development process used to build a system for performance evaluation.

The PEM proposed in this paper is designed to (1) evaluate the performance of the biometric recognition algorithm, and (2) to build a system that automatically evaluates performance and outputs the results in tandem with the biometric recognition system.

### *3.1. Structure of the PEM*

The PEM structure for the system that evaluates the biometric recognition algorithm is composed of a data preparation module, an execution model, and a result analysis module, as shown in Figure 1.

#### *3.1.1. Data preparation module*

The data preparation module prepares the biometric information used for performance evaluation, for which the development of a biometric information database and the design of the test criteria are the major elements to consider. As the biometric information database used affects the evaluation reliability of the performance evaluation system to a large extent, it should be considered a priority at the initial stage of system development. In addition, the biometric information used for performance evaluation should never be exposed, so that evaluation reliability can be improved [9].

Generally, algorithm performance evaluation of biometric recognition technology is conducted in such a way that the standard gallery is trained or registered, and is then compared with the test biometric information (probe) to be recognized, after which the similarities between the two sets of information are measured. At this time, algorithm performance varies according to the information generation environment or conditions. For example, if an expressionless front-view facial image is registered in the gallery, and a smiling facial image photographed at a 15-degree angle from the left is used as the probe, we can compare the strength of the different algorithm technologies in terms of facial expression and angle. In this paper, the "test criteria" refers to an item that could affect the performance evaluation result, and the performance evaluation system developer should design test criteria that are suitable for the objective of the evaluation. The test criteria selected by the PEM are limited to the classification criteria of the biometric information database.

#### *3.1.2. Execution module*

The execution module activates the biometric recognition system to be evaluated, and executes a performance evaluation. There are two methods for establishing an interface between the performance evaluation and the biometric recognition system. The first one consists in developing two systems as independently applied programs, while the second one consists in creating the biometric recognition algorithm as a component or library and then inserting it into the performance evaluation system. The former requires advance agreement between two systems with regards to the input/output file format, since the input data used for performance evaluation and the performance evaluation execution result data are generally transferred in a predefined form (generally, XML). Even though FRVT 2006 did not use the performance evaluation tool, participating companies submitted the biometric recognition system as an execution file, and the name of the input file used for evaluation and the output file that records the evaluation result were transferred as the program argument. For the component (or library)

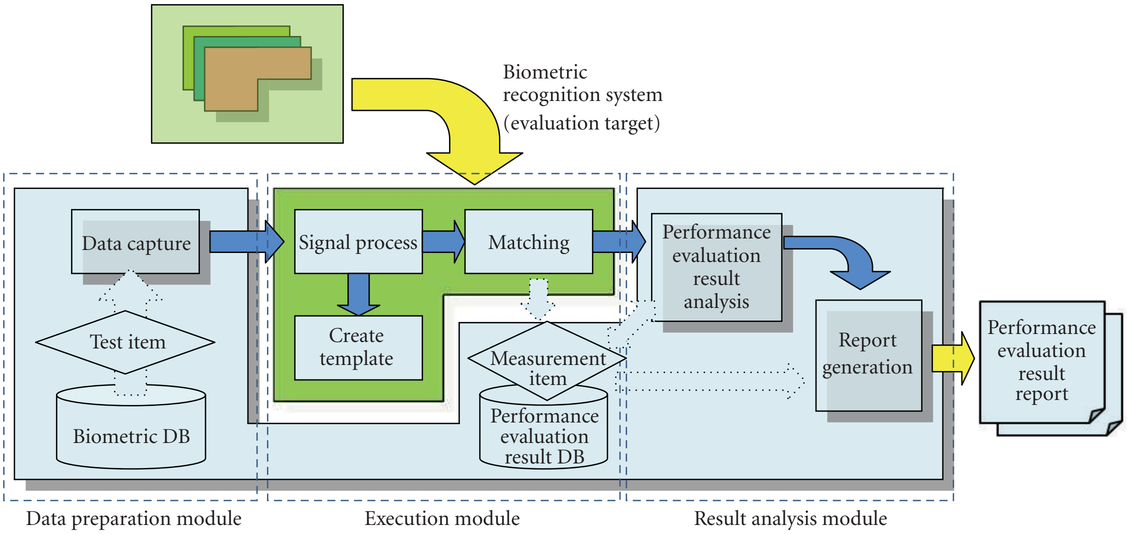


Figure 1: PEM Framework.

method, the standardized interface should be agreed upon in advance. The agreed interface should be as simple as possible, and compatibility with international standards is desirable. The related international standards include biometric application programming interface (BioAPI) 1.1 [10] and BioAPI 2.0.

#### 3.1.3. *Result analysis module*

The result analysis module performs a final analysis of recognition algorithm performance, using the result value obtained from the execution module. The performance of the specific algorithm can be expressed using several measurement criteria, and the appropriate measurement factor is decided upon depending on the objectives of the performance evaluation. Measurement factors can be broadly grouped into error rates and throughput rates. Error rates are basically derived from matching errors and sample acquisition errors, and the focus is on whether the algorithm is working properly and accurately. The throughput rate shows the number of users that the face recognition system can process in a given unit time. This throughput rate has significant meaning when performing the verification in a large image database [7].

### 3.2. *Formalization of PEM*

As biometric products are being used in establishing national infrastructure, a need for more effective and objective biometric performance test is on the increase. As examples are being announced such as FRVT, however, no objective and proved methodology is reported yet. In this paper, by presenting and formalizing biometric system performance test models, firstly, securing efficiency by eliminating unnecessary processes or factors that can occur when evaluating the performance of a biometric recognition system, and secondly, guaranteeing objectivity may be accomplished by generating credible test factors and processes for the performance test, which is completely dependent on heuristic methodology.

The performance test was executed by elevating the usability of presented models in this paper, and by formulating model-based tools for validation.

PEM is defined as a structure $\Gamma = (\mathrm{DB}, \mathrm{ATTR}, \mathrm{INT}, \mathrm{MET})$ with the following meanings.

(a) DB is a set of all images of the test database.

(b) ATTR is a set of pair $\langle \mathrm{pattr}, \mathrm{gattr}\rangle$ representing classification factors for probe and gallery set.

  (i) $\mathrm{pattr}, \mathrm{gattr} \subseteq \mathrm{fact}$ where fact is a set of factors influencing performance.

  (ii) fact = {age, sex, background, expression, pose, illumination,. . .}.

(c) INT is an interface of biometric recognition system being tested.

(d) MET is a set of performance metrics.

DB represents all facial images in the image database to be used for the purpose of face recognition performance evaluation. ATTR represents the factors that affect the performance test. The factors include age, sex, photographing time, expression, background, posture, lighting, and costume; and these factors are used when selecting probe and gallery image set. Probe image set is $\mathrm{PSET} = \boldsymbol{\sigma}_{\mathrm{pattr}}(\mathrm{DB})$, and gallery image set is $\mathrm{GSET} = \boldsymbol{\sigma}_{\mathrm{gattr}}(\mathrm{DB})$. $\boldsymbol{\sigma}_{\mathrm{condition}}$ is a function selecting elements that satisfy the condition. For example, in case of $\mathrm{ATTR} = \{\langle\{\mathrm{Man}, \mathrm{ExprSmile}\}, \{\mathrm{Front}, \mathrm{Normal}\}\rangle\}$ to perform the performance test for laughing men's faces, the probe image set used in the performance test is $\boldsymbol{\sigma}_{\mathrm{Man},\mathrm{ExprSmile}}(\mathrm{DB})$. Gallery image set is $\boldsymbol{\sigma}_{\mathrm{Front},\mathrm{Normal}}(\mathrm{DB})$.

When executing the performance test, all images of GSET will be matched to each image of PSET. That is, all image matching set MSET is a Cartesian product of PSET and GSET.

*(a) MSET* $= \{(\mathbf{x},\mathbf{y}) \mid \mathbf{x} \in$ *PSET and* $\mathbf{y} \in$ *GSET*$\}$

**INT** is the interface of face recognition module, an object of the performance test. Through this interface, performance-testing tools call the function of face recognition module. The interface should be as simple as possible, and it is desirable to be compatible with the international standards. INT should include the matching function $F_{\text{matching}}$, and this function outputs the matching results (accept or reject) by accepting one factor of image matching set.

*(b)* $\exists F_{matching} \in$ *INT, such that* $F_{matching}(a)$ *returns "accept" or "reject," where* $a \in$ *MSET*

The execution result of face recognition function for the performance evaluation is expressed as the results from calling $F_{\text{matching}}$ function using all element of MSET as factors, and this can be expressed as a two dimensional matrix, RMATRIX. That is, $\text{RMATRIX}_{i,j}$ element value is $F_{\text{matching}}(\text{PSET}_i, \text{GSET}_j)$. When the result value is accept, the outcome is 1, when the resulting value is reject, it is 0. For example, PSET $= \{p1, p2, p3\}$, and GSET $= \{g1, g2, g3\}$, MSET becomes $\{\langle p1,g1\rangle, \langle p1,g2\rangle, \langle p1,g3\rangle, \langle p2,g1\rangle, \langle p2,g2\rangle, \langle p2,g3\rangle, \langle p3,g1\rangle, \langle p3,g2\rangle, \langle p3,g3\rangle\}$. If the resulting values of applying each element of MSET to $F_{\text{matching}}$ is 1, 0, 0, 0, 1, 1, 0, 0, 1, consecutively, this can be expressed as the following 2-dimensional matrix:

$$\text{RMATRIX} = \begin{array}{c} \\ p1 \\ p2 \\ p3 \end{array} \begin{array}{c} \begin{array}{ccc} g1 & g2 & g3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}. \tag{1}$$

MET is a set of performance test measures such as fail-to-enroll rate (FTER), fail-to-acquire rate (FTAR), and false nonmatch rate (FNMR), false match rate (FMT). Such matching error-related metrics as FNMR, FMT can be calculated using RMATRIX:

$$\begin{aligned} &\text{FNMR} \\ &\quad = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\text{Same}(\text{PSET}_i,\text{GSET}_j) \wedge (\text{RMATRIX}_{i,j}=0)\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\text{Same}(\text{PSET}_i,\text{GSET}_j)}, \\ &\text{FMR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\text{Diff}(\text{PSET}_i,\text{GSET}_j) \wedge \text{RMATRIX}_{i,j}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\text{Diff}(\text{PSET}_i,\text{GSET}_j)}, \end{aligned} \tag{2}$$

where the following hold:

(i) $n$ is the size of PSET,
(ii) $m$ is the size of GSET,
(iii) $\text{Same}(a,b)$ is 1, if $a$ and $b$ are images of same person, otherwise $\text{Same}(a,b)$ is 0,
(iv) $\text{Diff}(a,b)$ is 0 if $a$ and $b$ are images of same person, otherwise, $\text{Diff}(a,b)$ is 1,
(v) $a \wedge b$ is 1, if $a$ and $b$ are 1, otherwise it is 0,
(vi) $a = b$ is 1, if values of $a$ and $b$ are same, otherwise it is 0.

### 3.3. Evaluation system development process

The following section describes how to build a performance evaluation system according to the PEM.

(1) Describe the objectives of developing and evaluating a performance evaluation system.
(2) Develop or select the biometric information database that will be used for performance evaluation.
(3) Design test criteria that fit into the evaluation objectives.
(4) Determine the type of interface to be used between the performance evaluation system and the biometric recognition system. If it runs as a standalone program, select the input/output file format; or, if it is linked by the component method, design the component interface.
(5) Select the measurement criteria that fit into the evaluation objectives.
(6) Implement the "data preparation module" which reads the biometric information according to the test criteria through the interface with the biometric information database.
(7) Implement the "execution module" which executes the recognition algorithm with the gallery/probe biometric information provided by the data preparation module. The execution result (degree of similarity) should be saved in the database containing the results of performance evaluation.
(8) Calculate the value of the measurement criteria by analyzing the similarity saved in the performance evaluation result database, and implement the "result analysis module" to generate a report on the results of the performance evaluation.

The test items and measurement criteria can be decided when performing the actual performance evaluation instead of building the performance evaluation system. In this case, select test items and measurement criteria that can be selected at steps 3 and 5 as described above, the test should be able to select the necessary items from among these when running the performance evaluation tool.

## 4. DESIGNING AND IMPLEMENTING THE PERFORMANCE EVALUATION TOOL

The performance evaluation tool was designed and implemented using the PEM proposed by this paper. The following section describes the contents and results by step, according to the evaluation system development process. The purpose of performance evaluation is to identify the technology level of the face recognition system through objective performance evaluation and certification, so as to encourage public trust in face recognition products and enhance their competitiveness.

### 4.1. Test criteria

The test criteria are designed in such way that those do not have to be selected when developing the evaluation system, enabling the tester to select it in the course of performing the actual performance evaluation. Basically, the test item lists all the classification criteria so that the tester can select from them separately, based on the condition of the gallery image set and the probe image set. The gallery image set and the probe image set, each of which is composed of several items, are referred to as the "test set." One performance evaluation project can generate several test sets, and each test set can generate a different result report. For instance, even though the collection of facial images to be registered in the face recognition system is white-front (normal) or purple-front (illum. yellow), the actual image acquired by the image acquisition device to identify the user can be normal, eye-closed, or tilted slightly to the left or right. The test set can be configured as in Figure 2. by assuming this kind of face recognition system.

If we assume that 10 face images exist per category in the above example, there are 20 facial images in the gallery set, and 40 facial images in the probe set. Therefore, the template generation function will be invoked 20 times for the gallery image, and image processing will be performed 40 times for the probe image, which means that 800 (20 × 40) comparison operations will be executed if performance evaluation is performed on the above test set. Among these comparison operations, image comparison will be performed 80 times for all the persons involved, because the image for a specific person appears twice in the gallery set and 4 times in the probe set, which results in an image comparison being performed 8 times for the same person, with 10 persons in total being compared. Therefore, 720 different image comparisons are made for the different persons, based on this system. Using this method, image comparison times can be estimated in advance, and the tester can estimate the time required for performance evaluation in advance, because the comparison times are calculated beforehand.

### 4.2. Selected BioAPI

The performance evaluation tool provides a standard interface for the face recognition system, and the examinee provides the face recognition module that satisfies this interface as the dynamic library. The performance evaluation tool is designed to enable the tester to change the face recognition module during the run time so that the tester can perform evaluation by changing the face recognition module without modifying the performance evaluation tool.

BioAPI 1.1 was applied for the compatibility with international standards, and only the minimum number of functions required for algorithm performance evaluation was selected, in order to reduce the examinee's development burden. Table 3 shows the selected BioAPI for the face recognition module.

### 4.3. Measurement criteria

The performance evaluation metrics used by FERET and FRVT, as well as the metrics proposed by JTC 1/SC 37/WG 5 Standard [7], were analyzed. Among these metrics, the criteria related with the technology evaluation of the face recognition system were chosen, as shown in Table 4.

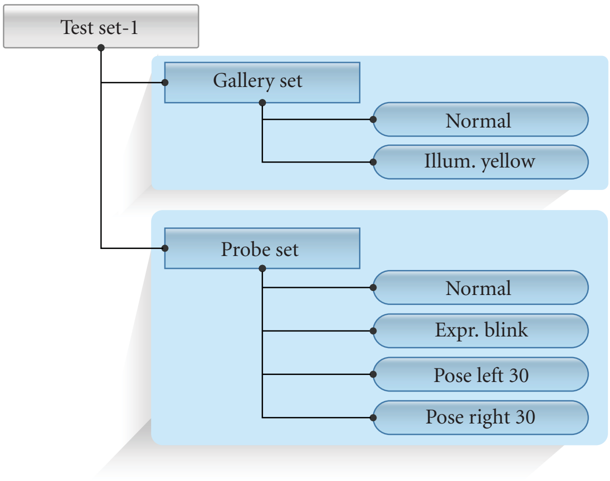


Figure 2: Example of setting the facial image test set.

### 4.4. Class diagram of metadata

Within the performance evaluation tool implemented in this paper, individual projects created for performance evaluation internally generate metadata in a specific structure in order to save the settings related to performance evaluation and performance evaluation results. The structure of these metadata is as illustrated in Figure 3.

#### 4.4.1. CMetaProject class

Whenever a new project is created, this class creates an instance in connection with the project. It maintains the path used to save the project name and the project itself, and the path to access a database in character strings. This information is related to the project configurations, and contains values established by the tester upon the project's creation. In addition, the list of face image data designed by the tester for performance evaluation receives the lists of CMetaTestSet class. A single project may have at least one group of face image data for its performance evaluation, and independently create a report based on these individual evaluation results. Therefore, the information exists in the extendible list format. Moreover, this class allows users to ascertain the number of total images to test (compare), and that of probes and gallery images with regards to such information.

#### 4.4.2. CMetaCategory class

This class contains data related to the subcategories of face images. Face images are influenced by the direction of the picture taken, the location of lighting, and posture, and so forth. According to these conditions, they are classified, and the tester may choose some of the classified images by using the performance test tool as the test subject group.

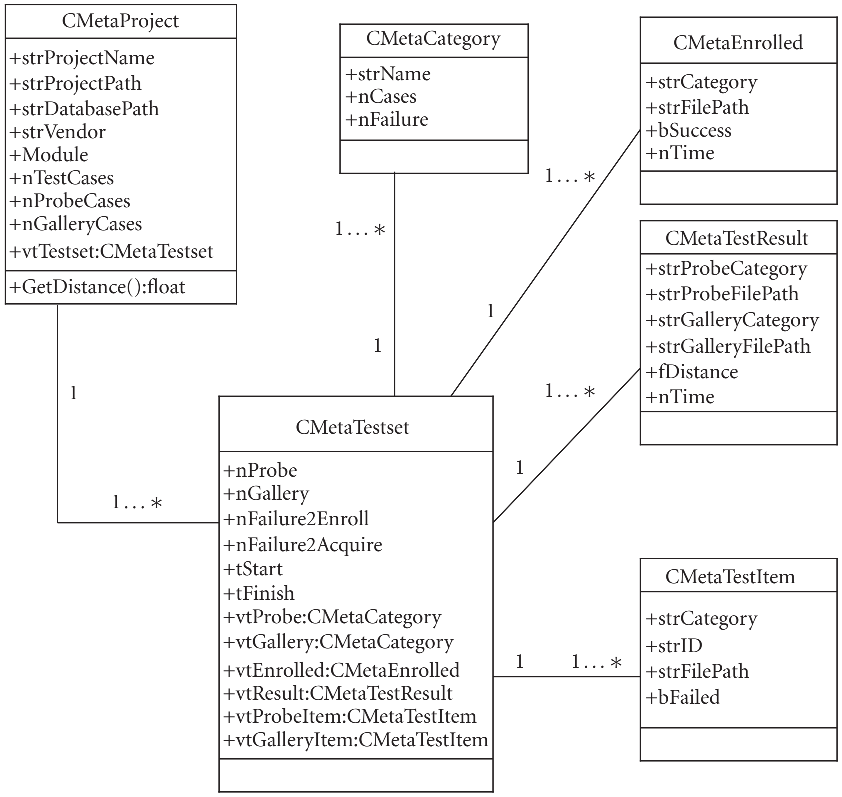


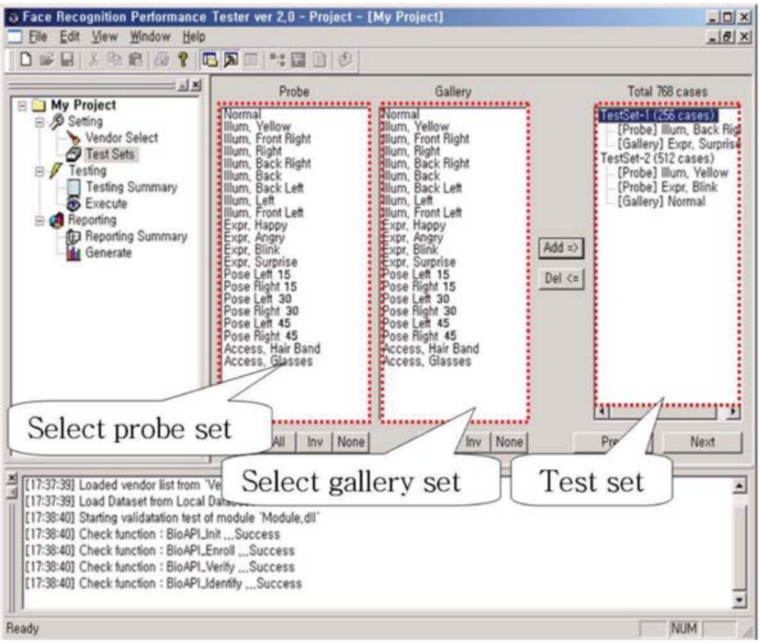
Figure 3: Example of setting the facial image test set.

Table 3: Selected BioAPI for face recognition module.

| Function name | Meaning |
|---|---|
| BioAPI_Init | initializes the face recognition module |
| BioAPI_Terminate | terminates the face recognition module |
| BioAPI_FreeBIRHandle | releases the BIR data memory |
| BioAPI_GetBIRFromHandle | returns the BIR data from the data handle |
| BioAPI_CreateTemplate | generates the data template from the gallery |
| BioAPI_Process | generates the BIR data by processing the probe image |
| BioAPI_VerifyMatch | returns the verification result by comparing two processed images |

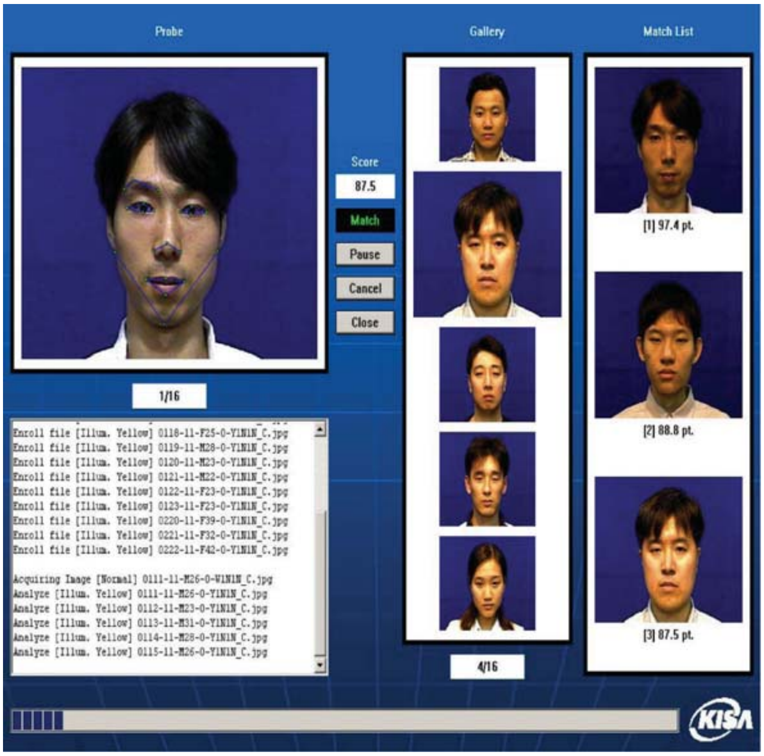
Table 4: Selected performance evaluation metric.

| Metric type | |
|---|---|
| Fail-to-enroll rate | enrollment throughput (no. of cases/s) |
| Fail-to-acquire rate | enrollment time (min, max, avg) |
| False reject rate (= fta + fnmr$^*$(1-fta)) | extraction throughput (no. of cases/s) |
| false accept rate (= fmr$^*$(1-fta)) | extraction time (min, max, avg) |
| false nonmatch rate | matching throughput (no. of cases/s) |
| False match rate | matching time (min, max, avg) |
| cmc curves | transaction throughput(no. of cases/s) |
| roc curves | transaction time (min, max, avg) |
| error equal rate | template size (min, max, avg) |



```
[17:37:39] Loaded vendor list from 'Ve
[17:37:39] Load Dataset from Local Data
[17:38:40] Starting validatation test of module 'Module.dll'
[17:38:40] Check function : BioAPI_Init ...Success
[17:38:40] Check function : BioAPI_Enroll ...Success
[17:38:40] Check function : BioAPI_Verify ...Success
[17:38:40] Check function : BioAPI_Identify ...Success
```

Figure 4: Selection window for face image probe set and gallery set.



```
Enroll file [Illum. Yellow] 0118-11-F25-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0119-11-M28-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0120-11-M23-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0121-11-M22-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0122-11-F23-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0123-11-F23-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0220-11-F39-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0221-11-F32-0-Y1N1N_C.jpg
Enroll file [Illum. Yellow] 0222-11-F42-0-Y1N1N_C.jpg

Acquiring Image [Normal] 0111-11-M26-0-W1N1N_C.jpg
Analyze [Illum. Yellow] 0111-11-M26-0-Y1N1N_C.jpg
Analyze [Illum. Yellow] 0112-11-M23-0-Y1N1N_C.jpg
Analyze [Illum. Yellow] 0113-11-M31-0-Y1N1N_C.jpg
Analyze [Illum. Yellow] 0114-11-M28-0-Y1N1N_C.jpg
Analyze [Illum. Yellow] 0115-11-M26-0-Y1N1N_C.jpg
```

Figure 5: Progress visualization when evaluating performance.

The chosen information is individually saved in the Cmeta-category class. These metadata contain variables, such as the category name of each criterion, the total number of images in the category, and the number of images failed to enroll (for gallery items) or acquire (for probe items).

### *4.4.3. CMetaTestItem class*

This class contains data pertaining to one face image. Each face image has a unique ID for the photograph target, its location, and the face image items. Additionally, it contains

Boolean variables in order to save whether each item failed to enroll (for gallery items) and acquire (for probe items) or not.

#### 4.4.4. CMetatTestResult class

This class stores the verification results of the comparison of one probe item with one gallery item in order to determine whether or not they come from an identical person. It contains each item's criteria, location, and ID information, along with variables, such as the similarity value and the comparison time created as a result of a comparison between two images.

#### 4.4.5. CMetaEnrolled class

This class saves the template data created so as to recognize a face from the image data for the system to execute the enrollment of a gallery item. It holds not only the item's criteria and location information but also a binary space to store template data, as well as data used to save the results of the template creation and the time required to create a template.

#### 4.4.6. CMetaTestset class

This class includes information about the face image probe group and gallery group in order to conduct the test. One project may have several test groups, and each test group individually saves the performance evaluation results. Metadata of the test groups bear the following information. Firstly, the class contains CMetaCategory which is the information holding the face image criteria as the list information for each of the probe and gallery groups. Namely, each of the probe and gallery groups may include a face image group with several different criteria. In addition, the class maintains the list of CMetaTestItem for each of the probe and gallery groups. This is not the criteria information but the metadata with individual image item information used for the test. It also contains the list of CMetaTestResult, containing the test results between a single probe item and a single gallery item.

Furthermore, where a system enrolls a gallery item, this class will contain the list of CMetaEnrolled classes in order to store template data that are created when each face recognition module generates its own template data. This is accomplished by using the face image data. As general data of such list data, the class contains variables to contain the test start time, end time, number of total probe items, number of total gallery items, number of total gallery items that failed to enroll, and number of total probe items that the system failed to acquire. We developed the six major modules and metadata classes that we examined above with the program for Windows, under Microsoft's Visual Studio development configuration. The face recognition modules, which are the subject of the performance evaluation, work in connection with the performance evaluation tool in the format of a dynamic link library.

### 4.5. Implementing data preparation/execution/result analysis module

The performance evaluation tool was developed as an application program running on Windows OS, and the face recognition module to be evaluated was implemented as a dynamic link library (DLL). The data preparation module that has the function of connecting with the biometric database and of setting the gallery and probe image set was implemented for use in the performance evaluation, as shown in Figure 4.

The face recognition module provided by the vendor was checked to verify that it provides the functions presented in Table 3, and the execution module that performs evaluation was implemented, using the functions of the selected face recognition module. A function that visually displays whether performance evaluation is progressing properly or not was included in the execution module, as shown in Figure 5.

Finally, the value of evaluation criteria is calculated by analyzing the similarity saved in the performance evaluation result database, and the result analysis module that generates the performance evaluation result report is implemented. This performance evaluation tool is equipped with a function that generates the evaluation result, as well as a function that issues the certificate for the face recognition module, depending on the evaluation result.

## 5. COMPARISON OF PERFORMANCE EVALUATION METHODS

Table 5 shows a comparison made between FERET and FRVT, which are the representative face recognition evaluation cases, and the evaluation method that uses the performance evaluation tool proposed by this paper.

Compared with performance evaluation programs such as FERET and FRVT, the performance evaluation tool proposed by this paper provides the following benefits.

(i) Disclosure of the face image database can be fundamentally prevented.

(ii) Development of a face recognition module that complies with international standards will be encouraged.

(iii) The performance evaluation target can be separated from the performance tester.

(iv) The evaluation cost can be reduced significantly, and individual evaluations can be performed for each vendor.

## 6. CONCLUSION

This paper proposed a PEM to evaluate the performance of biometric recognition systems. The proposed PEM is designed for compatibility with the related international standards, thereby contributing to the enhanced consistency and reliability of the performance evaluation tool that is developed according to this design. The proposed PEM is essential for the following reasons.

Table 5: Comparison between evaluation methods (FERET, FRVT, and the proposed performance evaluation tool).

| | FERET | FRVT | Proposed evaluation tool |
|---|---|---|---|
| Evaluation target | Face recognition system | Face recognition system<br>Scenario<br>Operation | Face recognition system |
| Face DB distribution method | DB distribution onsite (possibility of face DB disclosure) | DB distribution on-site (possibility of face DB disclosure)<br>FRVT 2006: Vendors provide the face recognition module. (No risk of disclosing face DB) | Vendors provide the face recognition module. (No risk of disclosing face DB) |
| Performance evaluator | Vendor staff | Vendor staff<br>FRVT 2006: Evaluator | Evaluator |
| Result analysis tool | Performance result analysis tool is required | Performance result analysis tool is required | No analysis tool is required |
| Result report generation | Manual work | Manual work | Automatically generated by the tool |
| Individual evaluation | Impossible | Impossible | Individual valuation by vendor |
| Evaluation costs | All related vendor staff should congregate in one place (expensive) | All related vendor staff should congregate in one place (expensive)<br>FRVT 2006: No meeting is required (low cost) | No meeting is required (low cost) |
| Responsibility of the vendor | I/O file format should be complied with | I/O file format should be complied with | Complies with recognition module interface |

(i) It represents a model and development method for the performance evaluation system.

(ii) It applies the related international standards to the performance evaluation system.

(iii) It enhances the consistency and reliability of the performance evaluation system.

(iv) It provides guidelines for the design and implementation of the performance evaluation system by formalizing the performance test process.

In addition, a performance evaluation tool capable of comparing and evaluating the performance of the commercialized facial recognition systems was designed and implemented, and an evaluation that executed 800 billion comparisons in 596 hours using the KFDB [8] was conducted. The certificate issuance criteria regarding the performance of the face recognition systems should be presented systematically, and a method should be prepared that can promote certification.

## REFERENCES

[1] P. J. Phillips, P. Rauss, and S. Der, "FERET (FacE REcognition Technology) Recognition Algorithm Development and Test Report," ARL-TR-995, U.S. Army Research Laboratory, 1996.

[2] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[4] D. M. Blackburn, J. M. Bone, and P. J. Phillips, "FRVT 2000 Evaluation Report," FRVT 2002 documents, February 2001.

[5] P. Grother, R. J. Micheals, and P. J. Phillips, "Face recognition vendor test 2002 performance metrics," in *Proceedings of the 4th International Conference on Audio- and Video-Based Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 937–945, Guildford, UK, 2003.

[6] P. J. Phillips and R. J. Michaels, "FRVT 2000: Evaluation Report," FRVT 2002 documents, March 2003.

[7] ISO/IEC JTC 1/SC 37, "Biometric performance testing and reporting—part 1: test principles and framework," ISO IS, 2006.

[8] B.-W. Hwang, H. Byun, M.-C. Roh, and S.-W. Lee, "Performance evaluation of face recognition algorithms on the Asian face database, KFDB," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 557–565, Guildford, UK, June 2003.

[9] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction to evaluating biometric systems," *Computer*, vol. 33, no. 2, pp. 56–63, 2000.

[10] "The BioAPI Consortium," BioAPI Specification Version 1.1, March 2001.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 513701, 10 pages
doi:10.1155/2008/513701

*Methodology Report*

# Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing

**Chon-Kit Kenneth Chan,[1] Arthur L. Hsu,[1] Sen-Lin Tang,[2] and Saman K. Halgamuge[1]**

[1] *Dynamic Systems & Control Group, Department of Mechanical Engineering, University of Melbourne, VIC 3010, Australia*
[2] *Research Center for Biodiversity, Academia Sinica, Taipei 115, Taiwan*

Correspondence should be addressed to Chon-Kit Kenneth Chan, c.chan22@pgrad.unimelb.edu.au

Received 31 August 2007; Accepted 18 November 2007

Recommended by Daniel Howard

Metagenomic projects using whole-genome shotgun (WGS) sequencing produces many unassembled DNA sequences and small contigs. The step of clustering these sequences, based on biological and molecular features, is called binning. A reported strategy for binning that combines oligonucleotide frequency and self-organising maps (SOM) shows high potential. We improve this strategy by identifying suitable training features, implementing a better clustering algorithm, and defining quantitative measures for assessing results. We investigated the suitability of each of di-, tri-, tetra-, and pentanucleotide frequencies. The results show that dinucleotide frequency is not a sufficiently strong signature for binning 10 kb long DNA sequences, compared to the other three. Furthermore, we observed that increased order of oligonucleotide frequency may deteriorate the assignment result in some cases, which indicates the possible existence of optimal species-specific oligonucleotide frequency. We replaced SOM with growing self-organising map (GSOM) where comparable results are obtained while gaining 7%–15% speed improvement.

## 1. INTRODUCTION

Metagenomics is an emerging area of genome research that allows culture-independent, functional, and sequence-based studies of microbial communities in environmental samples. Whole-genome shotgun (WGS) sequencing has been applied to most of the metagenomic projects [1–6]. These projects have unveiled remarkable information on microbial genomics and also brought an unprecedentedly comprehensive and clearer picture of microbial communities. In the WGS sequencing approach, random sampling of DNA fragments of all microbes that form a community in an environmental sample is performed. The individual DNA fragments are sequenced and then assembled into genomes by using computing techniques. However, a fundamental limit of WGS sequencing is that only the genomes of high-abundance species can be completely or near-completely assembled [7] due to the requirement of multiple overlapping fragments for a confident assembly. In one of the prominent metagenomic studies conducted by Venter et al. [2], about 1 Gb of DNA sequences has been successfully sequenced from Sargasso Sea samples. This study has clearly indicated the existence of far more diverse microbial communities than previously thought. Most of the environmental genomes sequenced to date contain only few high-abundance species but many low-abundance species in the communities that account for a large portion of the total genome size of an environmental sample. The presence of large amount of DNA fragments from the low-abundance species poses a problem for assembling genomes. In order to infer the biological functions of a microbial community from sequences, a process named "binning" is used to group these unassembled DNA sequence fragments and small contigs into biologically meaningful "bins," such as phylogenetic groups [8].

There are a number of tools currently available for the binning process. These include Chisel System [9, 10], MetaClust [4, 11], TETRA [12, 13], PhyloPythia [14], and the combination of oligonucleotide frequency and SOM [15]. The Chisel System helps binning the sequences according to the identification, characterisation, and comparative analysis of taxonomic and evolutionary variations of enzymes. The remaining above-listed tools use the method of analysing nucleotide composition of sequences that is considered to have the potential of working well for the binning process in WGS sequencing [8]. MetaClust computes different DNA signatures followed by the use of a clustering algorithm to assign sequences into bins. TETRA bins the species-specific sequences by the use of tetranucleotide-derived $z$-score correlations. PhyloPythia uses a supervised-learning approach where it trains a multiclass support vector machine (SVM) classifier using all the known genome sequences in the existing database then assigns the unknown environmental sequences to the closest clade in the selected taxonomic level. This method has been demonstrated to be able to classify most DNA sequence fragments with high accuracy. However, considering the current amount of known genomes which is far less than 1% of the entire microbial genomes [7], it is reasonable to assume that the currently available training data is insufficient to represent all the extremely diverse microbial genomes for supervised-learning methods. Unsupervised-learning may provide the answer to this problem. The combination of oligonucleotide frequency and the well-known unsupervised learning method self-organising map (SOM) was used by Abe et al. [15] to explore genome signatures. They used the di-, tri-, and tetranucleotide frequencies as the training features of SOM to cluster the 1 kb and 10 kb DNA sequence fragments derived from 65 bacteria and 6 eukaryotes. Clear species-specific separations of sequences were obtained in the 10 kb fragment tests. Their results showed that the combination of oligonucleotide frequency and SOM can be used as a powerful tool to cluster or bin the DNA sequence fragments after WGS sequencing.

In order to successfully bin the DNA sequence fragments, using an appropriate genome signature as the training feature is important. In recent years, researchers have found that, due to oligonucleotide frequency bias in various prokaryotic genomes, the oligonucleotide frequency can be used as a possible genome signature. The di-, tri-, and tetranucleotide frequencies, which are the frequencies using two, three, and four nucleotides respectively, have been well studied. Karlin et al. [16, 17] has shown the compositional bias of the di- and tetranucleotide contents of 15 prokaryotic genomes. Weinel et al. [18] found that 80% of *Pseudomonas putida* KT2440 genome have a similar bias in GC contents and di- and tetranucleotide contents. Teeling et al. [12] showed that the tetranucleotide frequency has a higher discriminatory power than GC content and used it for the assignment of genomic fragments to the taxonomic group. In addition, Sandberg et al. [19] employed a Bayesian approach to classify the short sequences and found that the classification accuracy increases with a higher-order oligonucleotide frequency. Above-mentioned papers provide evidences that there is a trend of using oligonucleotide frequency as prokaryotic genome signature, rather than the GC content. Thus, high-order oligonucleotide frequency may also be an appropriate training feature for binning DNA sequence fragments by unsupervised clustering methods.

Since the combination of oligonucleotide frequency and SOM appears as a promising binning strategy that can be further explored, we focus in this paper on improving the training features and the clustering algorithm. In Abe et al.'s work [15], there was no systematic way of comparing the quality of the SOM results. We tested the traditional clustering evaluation measures (recall, precision, and F-measure) and discovered the inadequacy of using them for examining the similarity of phylogenetic levels. Therefore, we introduce a method to quantitatively measure and assess the results of clustering DNA sequence fragments from a collection of species. In the investigation of evaluating suitable training features, we attempt to compare results for the di-, tri-, and tetranucleotide frequencies as well as the pentanucleotide frequency (the frequency usage of five nucleotides) to test if higher-order oligonucleotide frequency yields better binning of DNA sequence fragments. We also study the effectiveness and efficiency of the combination of oligonucleotide frequency and SOM by employing alternative clustering algorithms. We compare SOM with a variant of it called growing self-organising map (GSOM), which has been successfully applied in several different applications [20–25] including microarray clustering [26]. These comparisons allow us to suggest a better compositional binning strategy for WGS sequencing using the method of combining oligonucleotide frequency and SOM-based clustering algorithm.

This paper is organized as follows: Section 2.1 gives a brief introduction to the SOM and GSOM clustering algorithms; Section 2.2 proposes a method of measuring inseparable species when DNA sequence fragments are clustered; Section 2.3 describes the procedures of preparing the three datasets used in this paper, and the data preprocessing step for preparing the input vectors; and Section 2.4 shows the details of the algorithm settings and the experiment set up for repeatability of the experiments. Section 3 presents the results of comparing the four orders of oligonucleotide frequencies and the comparison between SOM and GSOM; Finally, Section 4 gives the discussion, conclusion, and future work.

## 2. METHODS

### 2.1. Growing self-organising map

Growing self-organising map (GSOM) [27, 28] is an extension of self-organising map (SOM) [29]. GSOM is a dynamic SOM which overcomes the weakness of a static map structure of SOM. Both SOM and GSOM are used for clustering high-dimensional data. This is achieved by projecting the high-dimensional data onto a two- or three-dimensional feature map with lattice structure where every point of interest in the lattice represents a neuron or a node in the map. The mapping preserves the data topology, so that similar samples can be found close to each other on the 2D/3D feature map.

The SOM training consists of three phases: initialisation phase, ordering phase, and fine-tuning phase. The initialisation is crucial to achieve a quality-clustering result. The following parameters are determined in this phase:

(i) the map topology (either rectangular or hexagonal);
(ii) the number of nodes which is the resolution of the map;

(iii) the weight vector initialization of nodes;
(iv) the width/height (or aspect) ratio of the map.

The user determines the first two parameters and generally principle components analysis (PCA) is used for setting the last two parameters. The weight vectors are initialised by the first two principle vectors of the inputs and the aspect ratio of the map is determined based on the ratio of magnitudes of the first two principle components. In the ordering and fine-tuning phases, each input is presented to the map and the best matching unit or "winner," which has the smallest Euclidean distance to the presented input, is identified. The weight vector of the winner and its neighbouring nodes are updated by

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha \times h \times [\mathbf{x}(k) - \mathbf{w}(t)], \tag{1}$$

where $\mathbf{w}(t)$ is the weight vector of the node at time $t$, $\mathbf{x}(k)$ is the $k$th input vector ($\mathbf{w}, \mathbf{x} \in \Re^D$ where $D$ is the dimensionality of data), $\alpha$ is the learning rate and $h$ is the neighbourhood kernel function.

GSOM employs the same weight adaptation and neighbourhood kernel learning as SOM, but has a global parameter of growth named Growth Threshold (GT) that controls the resolution of the map. The Growth Threshold is defined as

$$\mathrm{GT} = -D \times \ln(SF), \tag{2}$$

where $SF \in [0, 1]$ is the user defined spread factor with 0 representing minimum growth (coarsest resolution) and 1 representing maximum growth (finest resolution).

There are three phases in the GSOM training: initialisation phase, growing phase, and smoothing phase. In the initialisation phase, the GSOM is initialised with a minimum single "lattice grid" depending on whether the rectangular or hexagonal topology is chosen. Due to the small number of nodes in the beginning of training, the weight vector initialisation has less effect on the clustering quality and these weights will be corrected quickly in the growing phase. During the growing phase, every node has an accumulated error counter and the counter of the winner ($E_{\text{winner}}$) is updated by

$$E_{\text{winner}}(t+1) = E_{\text{winner}}(t) + ||\mathbf{x}(k) - \mathbf{w}_{\text{winner}}(t)||. \tag{3}$$

If the winner is at the boundary of the current map and $E_{\text{winner}}$ exceeds $GT$, new nodes will be added to the surrounding vacant slots of the winner. In the case when $E_{\text{winner}}$ exceeds GT and the winner is not a boundary node, $E_{\text{winner}}$ is evenly distributed outwards to the winner's neighbouring nodes. The smoothing phase is for fine-tuning the weights of nodes and no new node will be added to the map.

The major advantages of GSOM over SOM are summarised as follows.

(i) The shape of GSOM represents the hidden data structure better than SOM that leads to better identifiable clusters.
(ii) New nodes are added to the necessary regions while keeping the order of nodes. Therefore, neither PCA nor ordering phase is required in the training.
(iii) Fewer nodes at the beginning of the training leads to the speed improvement.

### 2.2. Quality measurement of the clustering performance in the mixing region

In our preliminary test, the well-known F-measure [30], which computes both recall and precision into a single index from a contingency table, was used to evaluate the clustering results. However, after examining the cluster contents, it is apparent that, for binning applications, F-measure does not provide sufficient insight and description of ambiguities in terms of phylogenetic relationships (refer to Section 3). More specifically, one would expect phylogenetically-close groups as highly likely to be ambiguous, but F-measure does not account for such likelihoods. Therefore, we propose an alternative clustering evaluation measure specifically for this application.

When an SOM or GSOM is used to group species fragments into clusters on a 2D/3D map, it is often inevitable that regions with overlapping clusters (mixing regions where a neuron represents DNA sequence fragments from more than one species) will exist. To evaluate a clustering algorithm's ability to group DNA sequence fragments into species-specific or "pure" clusters, we define two criteria that measure the clustering quality in the mixing region: intensity of mix (IoM) and level of mix (LoM), where the former measures the percentage of mixing and the later indicates the taxonomic level of ambiguity for a given pair of clusters.

The IoM is evaluated based on the concept of mixed pair described below. Let $A$ and $B$ be sets of vectors belonging to species $A$ and $B$, respectively, and $n(X)$ is the number of elements in set $X$. If $A$ and $B$ is a mixed pair, then the percentage of $A$ in the mixing region of the two classes is $n(A \cap B \mid A)/n(A)$ and the percentage of $B$ is $n(A \cap B \mid B)/n(B)$. As illustrated in Figure 1, 11.6% of $A$ sequences is mixed with $B$ sequences in the $B$ cluster and the complementary mix indicates that 10.6% of $B$ sequences is mixed with $A$ sequences in the $A$ cluster. The same concept of mixed pair applies for $B$ and $C$. Therefore, there are two mixed pairs in Figure 1, one is $A$ and $B$ and the other is $B$ and $C$. For $k$ number of species, there can be up to $k(k-1)/2$ mixed pairs. Additionally, a pair of clusters is only considered to be truly mixed when both clusters are heavily overlapped. Thus, as in Figure 1, when $n((B \cap C \mid B)/n(B)) >$ THRESHOLD (TH) but $n(B \cap C \mid C)/n(C) <$ TH, it indicates that only a small number of outliers of one species ($C$) is mixed with the other species ($B$). Therefore, this mixed pair is not considered as truly mixed. We use TH = 5% for the threshold of being truly mixed meaning that, statistically, we have a nonmixing confidence of 95%. The IoM measures the amount of mixing sequences and it is nonlinearly categorised into five levels: low (L) 5%–10%, medium low (ML) 10%–20%, medium (M) 20%–40%, medium high (MH) 40%–60%, and high (H) 60%–100%. For example, the IoM is ML for the truly mixed pair $A$ and $B$ in Figure 1.

To evaluate clustering results of species, we use LoM to describe the taxonomic level of the mixed species. For example, as in Figure 1, *Bacillus subtilis* is classified in Kingdom *Bacteria* and Phylum *Firmicutes. Acinetobacter* is classified in Kingdom *Bacteria* but Phylum *Proteobacteria.* Then the two species are mixed at the Phylum level. Because of the

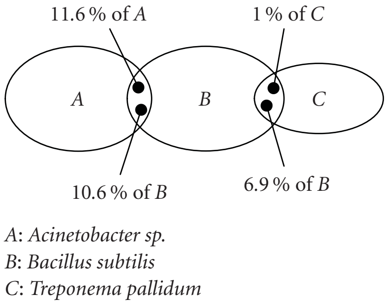


Figure 1: Concept of mixed pair: the mixed pair between $A$ and $B$ is truly mixed (IoM = ML and LoM = Phylum). The mixed pair between $B$ and $C$ is not truly mixed because $n(B \cap C \mid C)/n(C) < 5\%$.

evolution of organisms, nucleotide composition of genomes belonging to the same lower taxonomic levels can be very similar. Clustering organisms at higher level of taxonomy should be easier than at lower level of taxonomy. Therefore, if truly mixed pair occurs, lower LoM (e.g., Species) is more acceptable and more desirable than higher LoM (e.g., Kingdom).

In summary, the proposed two measures are defined as

$$\begin{aligned}&\text{IoM} \in \{\text{L}, \text{ML}, \text{M}, \text{MH}, \text{H}\},\\ &\text{LoM} \in \{\text{Species}, \text{Genus}, \text{Family}, \text{Order}, \text{Class},\\ &\qquad\qquad \text{Phylum}, \text{Kingdom}\}.\end{aligned} \tag{4}$$

The two proposed measures, IoM and LoM, are only defined for truly mixed pairs to evaluate the clustering quality in the mixing regions of a map by the following steps.

(i) Find truly mixed pairs for all pairs of species where if $n(X \cap Y \mid Y)/n(Y) \geq \text{TH}$ and $n(X \cap Y \mid X)/n(X) \geq \text{TH}$, then $X$ and $Y$ is a truly mixed pair.

(ii) If $X$ and $Y$ are truly mixed, determine IoM according to $\min\{n(X \cap Y \mid Y)/n(Y), n(X \cap Y \mid X)/n(X)\}$.

(iii) Identify LoM of $X$ and $Y$.

Clustering results can now be assessed based on three criteria: number of truly mixed pairs, IoM, and LoM. However, which criterion should have higher priority may vary between applications. Therefore, in our assessment, one result is better than another only when it is superior on at least two of the three measures.

### 2.3. Dataset preparation and data preprocessing

The NCBI database (http://www.ncbi.nlm.nih.gov) contains 370 completed microbial genomes in early 2006, which includes 28 Archaea and 342 Bacteria. As the investigation seeks to cluster sequences of species, genomes are filtered so that there is no duplicating species. One strain was arbitrarily chosen if the same species genome contains more than one strain. After this process, there were 283 genomes remaining. Considering the available computing resources and the algorithm comparison focus of this paper, two artificial sets of prokaryotic DNA sequences (each of 10 different species out of the 283 species) were randomly sampled from the NCBI database.

In addition, three simulated metagenomic datasets were created by Mavromatis et al. [31] to facilitate benchmarking of metagenomic data processing methods, which include, but not limited to, binning methods. The three datasets vary in relative abundance and number of species that represent different complexity levels of real-world microbial communities. The sequence fragments in the simulated datasets were assembled using three commonly used sequence assembling programs: JAZZ, Arachne, and Phrap at U.S. Department of Energy (Wash, USA), DOE Joint Genome Institute (Calif, USA). In this paper, we tested one of the three simulated datasets named simMC and was assembled by Phrap (http://www.phrap.com). For simplicity, this dataset will be represented as simMC_Phrap throughout the paper.

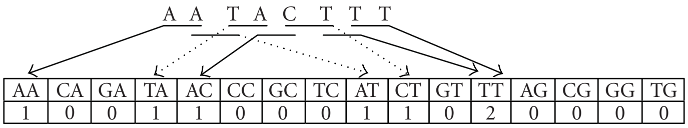
The taxonomic distributions of the three sets of species are displayed graphically in Figure 3. Each letter represents a single species and species within a single rectangle have the same taxonomy at the specific level. The names of the species can be found in Section 1 of the supplementary material which consists of 6 sections showing the species names in Section 1, clustering evaluation methods in Sections 2 and 3, and the labelled cluster maps in Section 4 to 6 (http://www.mame.mu.oz.au/~ckkc/Binning). The numbers below the taxonomic levels in Figure 3 indicate the maximum possible number of mixed pairs at that taxonomic level. For example, in Figure 3(a), the maximum number of mixed pairs at taxonomic level of *Class* is 12, which consists a mixed pair each from (a,j) and (c,e) and 5 pairs each from (c,{b,d,g,h,i}) and (e,{b,d,g,h,i}).

In the experiments of Abe et al. [15] that attempts to separate 1 kb and 10 kb DNA sequence fragments of 65 bacteria genomes containing 54 different species, it is visually shown that 1 kb DNA sequence fragments do not carry enough discriminatory information and hence could not completely separate the fragments into species-specific groups. Therefore, a sequence fragment length of 10 kb is used for the analysis in the two artificial datasets to ensure appropriate separation of species-specific groups. Sequences used in the two artificial datasets are produced from complete genome sequences to simulate the environment of WGS sequencing. Such a complete genome sequence is segmented into 10 kb nonoverlapping fragments. A sliding window with the size $n$ is used for counting the oligonucleotide frequency for each of the fragments in which $n$ is the nucleotide length. For example, the dinucleotide frequency ($n = 2$) for a short sequence "AATACTTT" is shown graphically in Figure 2. The oligonucleotide frequency count for each of the fragments yields a single input vector for clustering. The input vectors to the clustering algorithm will have $4^n$ dimensions.

Whereas, the simMC_Phrap was preprocessed by extracting all sequences with contig length $\geq$ 8 kb. The oligonucleotide frequency count is applied to these sequences to generate the input vectors for clustering. Finally, each input vector is normalised by the sequence length.

### 2.4. Algorithms parameters and experiment details

In order to avoid the algorithm implementation bias, an in-house clustering program was developed consisting a



| AA | CA | GA | TA | AC | CC | GC | TC | AT | CT | GT | TT | AG | CG | GG | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 |

FIGURE 2: Dinucleotide frequency counting for the short sequence "AATACTTT."

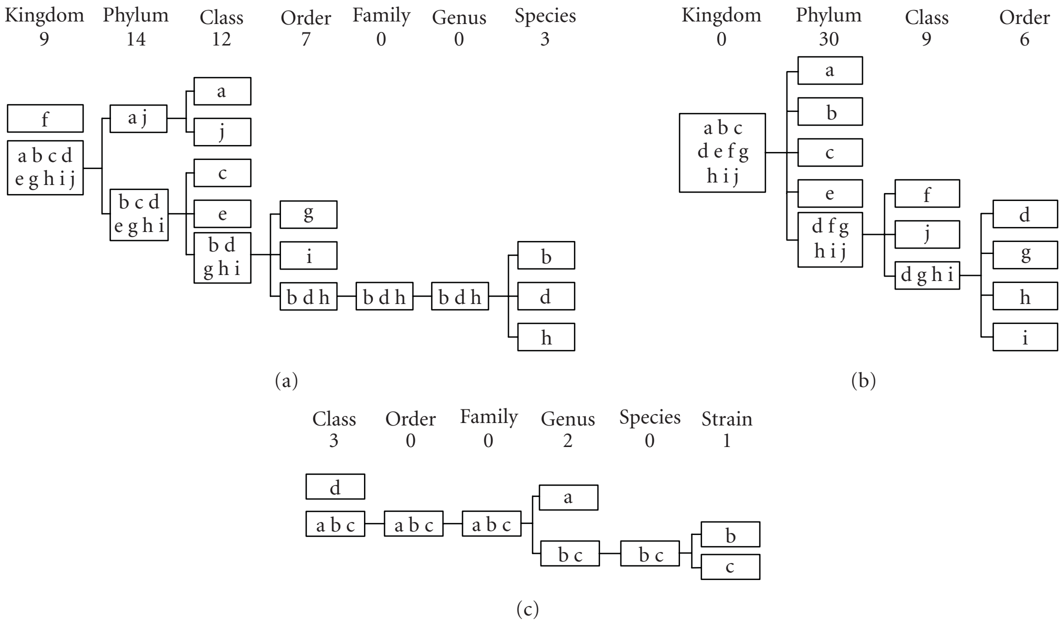


FIGURE 3: The taxonomy distribution of the 10 species in (a) Set 1, (b) Set 2, and the 4 species in (c) simMC_Phrap. Each letter represents a single species. The numbers below the taxonomic levels indicate the maximum number of mixed pairs at that taxonomic level. For example, in (a), the maximum number of mixed pairs at taxonomic level of Class is 12, which consists (a,j), (c,e), (c,{b,d,g,h,i}), and (e,{b,d,g,h,i}) mixed pairs.

TABLE 1: Training parameters used for the SOM and GSOM training.

| Training parameter | Phase 2 | Phase 3 |
|---|---|---|
| Learning length | 15 epochs | 70 epochs |
| Learning rate | 0.1 | 0.05 |
| Neighbourhood size | 3 | 1 |

data preprocessor, both SOM and GSOM algorithms, and a graphical neuron-map output. The program is written in C# with a Windows form interface. This program can be requested from the author for academic use. Since the hexagonal topology has better topology preservation [26, 29], it is used in both SOM and GSOM. In addition, the tests were conducted on a Pentium 4 3.2 GHz desktop PC running Windows XP and the same parameter settings are used in both algorithms for a fair computational speed comparison (as listed in Table 1).

To compare results from the 4 orders of oligonucleotide frequencies, we obtain similar map resolution (number of nodes) for both algorithms and for all nucleotide frequencies. Since the GSOM algorithm automatically determines the number of nodes, it can be used to determine the total number of nodes of SOM. This can be achieved by training GSOM with a specified resolution (we used SF = 0.4) for the dinucleotide frequency then the final number of nodes in GSOM is used to set the number of nodes in SOM, as well as determine the SF for other nucleotide frequencies. Using this scheme, we set SF = 0.4 for dinucleotide frequency (for a higher-resolution map) and experimentally determined that SF = 0.6 for trinucleotide frequency, SF = 0.8 for tetranucleotide frequency and SF = 0.9 for pentanucleotide frequency will result in similar map resolution. The SOM also requires setting the aspect ratio of the map and initializing the weight vectors. These two parameters are set by using PCA. The schematic diagram of this approach is shown in Figure 4.

## 3. RESULTS

The proposed binning method is tested on two artificial datasets and a simulated metagenomic dataset (simMC_Phrap) which was created and published for benchmarking the metagenomic data processing methods. The two artificial datasets of prokaryotic DNA sequences (each of 10 different species out of the 283 species) were randomly sampled from the NCBI database. Each set of the genome

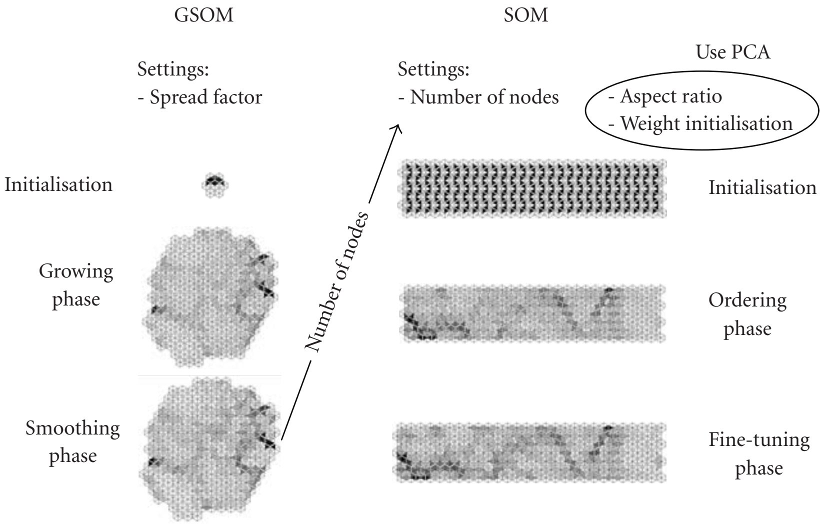


Figure 4: Illustration of method used to compare SOM and GSOM.

Table 2: The evaluation of clustering results using F-measure.

| | Set 1 | | Set 2 | | simMC_Phrap | |
|---|---|---|---|---|---|---|
| | SOM | GSOM | SOM | GSOM | SOM | GSOM |
| Di | 0.95 | 0.95 | 0.94 | 0.94 | 0.92 | 0.91 |
| Tri | 0.97 | 0.97 | 0.98 | 0.98 | 0.90 | 0.90 |
| Tetra | 0.97 | 0.97 | 0.98 | 0.98 | 0.92 | 0.90 |
| Penta | 0.97 | 0.97 | 0.99 | 0.99 | 0.89 | 0.89 |

sequences was preprocessed to obtain the 4 orders of oligonucleotide frequencies (di-, tri-, tetra-, and pentanucleotide frequencies) to form 4 datasets. This data preprocessing involves segmenting each genome sequence in the set into 10 kb lengths then produce input vectors by calculating the specific oligonucleotide frequency. After preprocessing, each of the 4 datasets from species Set 1 contains 4,145 input vectors. Whereas, each of the 4 datasets from Set 2 contains 2,398 input vectors. The simMC_Phrap was preprocessed by extracting all sequences with contig length ≥ 8 kb then obtaining the 4 orders of oligonucleotide frequencies to form 4 datasets. The produced input vectors were normalised by the sequence length. After preprocessing, each of the 4 datasets contains 401 input vectors. The details of preparing these three datasets can be found in Section 2.

To evaluate the clustering performance, we used well-known clustering evaluation measure F-measure. The results for the three datasets are shown in Table 2. A summary of F-measure calculation can be found in Section 2 of the supplementary material.

From these results, we can observe that F-measure does not distinguish the clustering quality clearly enough for this species separation application. For example, in the results of Set 1 using GSOM, F-measure equals 0.97 for the tri-, tetra-, and pentanucleotide frequencies. However, by using the proposed evaluation, more details of the ambiguities can be seen. It shows that there are only two mixed pairs with low taxonomic level in the mixing region when using the pentanucleotide frequency. However, there are four mixed pairs with two of them having higher taxonomic levels when using the tri- and tetranucleotide frequencies (as shown in Table 3). This suggests that the pentanucleotide frequency provides a higher level of phylogenetic resolution, which cannot be detected via F-measure because the numbers of incorrectly assigned sequence fragments are similar for these three nucleotide frequencies.

We use two approaches to evaluate the performance of clustering DNA sequence fragments of species. The first approach is to observe the cluster formation of species sequences to verify the cluster formation similar to the method used by Abe et al. [15]. The second approach is to compare the LoM and IoM in the mixing region. A simple example is given in Section 3 of the supplementary material. It highlights the difference between the calculation of F-measure and IoM.

After the training is completed, for display purpose, we use the label information of the input data to display the labelled cluster map. The labelled cluster maps from the training of SOM and GSOM for the pentanucleotide frequency of species Set 1 are shown in Figure 5. The following points can be observed from this labelled cluster maps.

(i) All species are clearly clustered and are marked in the figures.

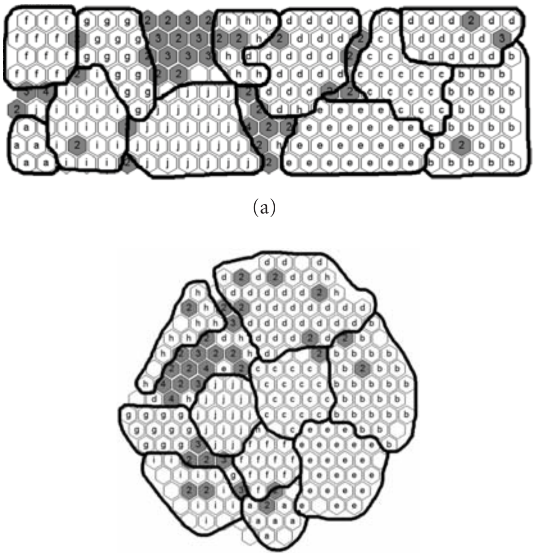
(ii) The nodes that contain more than 2 species (which are coloured in grey) are mostly located at the border of the clusters.

(iii) Species “d” is clustered as one group in GSOM, but separated by species “c” into two groups in SOM.

These observations show that the GSOM have better cluster formation in terms of cluster identification than the SOM due to the flexibility of feature map shape. The labelled cluster maps for other datasets also show a clear cluster

Table 3: Training results in the mixing regions for species Set 1.

| Algorithm | SOM | | | | GSOM | | | |
|---|---|---|---|---|---|---|---|---|
| Nucleotide Freq. | Di | Tri | Tetra | Penta | Di | Tri | Tetra | Penta |
| Kingdom | — | — | — | — | — | — | — | — |
| Phylum | — | — | — | — | — | — | — | — |
| Class | ML, ML, L | — | — | — | ML, ML, L | — | — | — |
| Order | ML, ML | ML, L | ML | L | M, L | ML, L | L, L | — |
| Family | — | — | — | — | — | — | — | — |
| Genus | — | — | — | — | — | — | — | — |
| Species | M, L | M, L | ML | ML | M, L | M, L | ML, L | ML, L |


(a)

(b)

Figure 5: The labelled cluster maps for clustering species Set 1 by (a) SOM, (b) GSOM with the pentanucleotide frequency. Each hexagon represents a single node. If a node contains input samples from only a single species, it is displayed with a letter that uniquely identifies the species. Grey colour nodes correspond to two or more species in the node and the number of species is displayed on the node. A node without label means that there is no input sample "hits."

formation and can be found in Sections 4, 5, and 6 of the supplementary material.

We also interpret the clustering results of the species mixing regions by summarising the IoM and LoM in Tables 3, 4, and 5 for Set 1, Set 2, and simMC_Phrap, respectively.

By comparing the results of the four orders of oligonucleotide frequencies for Set 1, the number of truly mixed pairs for dataset that uses the dinucleotide frequency is almost twice the number of truly mixed pairs of the higher-order oligonucleotide frequencies as shown in Table 3. In addition, the LoM is also high for the dinucleotide frequency. Similarly in Table 4, there are three to four truly mixed pairs when using the dinucleotide frequency, but no more than one truly mixed pair when higher-order oligonucleotide frequencies are used. All four truly mixed pairs are of a very high LoM at the Phylum level. Since there are only very few species and number of sequence fragments in simMC_Phrap, the difference of using different nucleotide frequencies is not obvious. Nevertheless, the results of our two artificial sets indicate that dinucleotide frequency is not a strong signature for clustering the 10 kb fragments of species.

Furthermore, from Set 1, we can see that IoM and LoM tend to decrease as the order of oligonucleotide frequency increases. One would naturally suspect that higher-order oligonucleotide frequencies may carry more information so they can be used to achieve better clustering results. However, it is not the case for Set 2 and simMC_Phrap. In Set 2, there is a truly mixed pair in the tetranucleotide frequency but no mixed pair in the tri- and pentanucleotide frequencies when using SOM. For the GSOM algorithm, a truly mixed pair appears in the pentanucleotide frequency but not in the tri- and tetranucleotide frequencies. A detailed examination shows that they are the same pair, *Acinetobacter* sp.ADP1 and *Bacillus subtilis* subsp. *subtilis* str. 168, in both cases. Both of the IoM is low indicating a much better result than the dinucleotide frequency. As in Table 5 for simMC_Phrap, GSOM performs well for all four orders of nucleotide frequencies, whereas SOM shows inconsistent clustering quality for different nucleotide frequencies. There are two mixed pairs in the di- and pentanucleotide frequencies but only one mixed pair in the other two nucleotide frequencies. The mixed pair in the tetranucleotide frequency has the lowest IoM (IoM = M). These results also do not support the hypothesis that higher-order oligonucleotide frequencies are better clustering features. Therefore, we can only conclude that higher-order oligonucleotide frequencies are better features for clustering the species than using dinucleotide frequency. However, the optimal oligonucleotide frequency may vary in different species.

In terms of clustering quality, both SOM and GSOM have similar results. However, besides the mixing quality comparison, we also compare the training speed of them. The speed comparisons for the first two training phases and the overall training time of both algorithms for all 12 datasets are shown in Tables 6 and 7, respectively.

Comparing the time taken for SOM and GSOM to finish the first two training phases (as in Table 6), GSOM has more than 37% speed improvement than SOM. This speed improvement can be explained by considering the initial formation of the map structure. As discussed in Section 2, PCA

Table 4: Training results in the mixing regions for species Set 2.

| Algorithm | SOM | | | | GSOM | | | |
|---|---|---|---|---|---|---|---|---|
| Nucleotide Freq. | Di | Tri | Tetra | Penta | Di | Tri | Tetra | Penta |
| Kingdom | — | — | — | — | — | — | — | — |
| Phylum | MH, ML, ML | — | L | — | H, ML, L, L | — | — | L |
| Class | — | — | — | — | — | — | — | — |
| Order | — | — | — | — | — | — | — | — |
| Family | — | — | — | — | — | — | — | — |
| Genus | — | — | — | — | — | — | — | — |
| Species | — | — | — | — | — | — | — | — |

Table 5: Training results in the mixing regions for the contigs ≥ 8 kb from simMC_Phrap.

| Algorithm | SOM | | | | GSOM | | | |
|---|---|---|---|---|---|---|---|---|
| Nucleotide Freq. | Di | Tri | Tetra | Penta | Di | Tri | Tetra | Penta |
| Kingdom | — | — | — | — | — | — | — | — |
| Phylum | — | — | — | — | — | — | — | — |
| Class | — | — | — | — | — | — | — | — |
| Order | — | — | — | — | — | — | — | — |
| Family | — | — | — | — | — | — | — | — |
| Genus | MH | MH | M | MH | MH | MH | MH | MH |
| Species | — | — | — | — | — | — | — | — |
| Strain | L | — | — | L | — | — | — | — |

is used to initialise SOM. However, it increases the computational cost exponentially when the data dimension and size increases. Therefore, even though time consumed for PCA initialisation in this experiment is negligible, it is expected to be significant in large-scale metagenomic analysis. In addition, SOM starts with all nodes fully covering the whole input space then adjusts the weights of all nodes to represent the input data better. On the other hand, GSOM starts with the minimum number of nodes and grows more nodes in the required direction to get an abstract representation of the input data while still correcting the weights of existing nodes. At the end of the second phase, both algorithms will be roughly representing the abstraction of the data. However, GSOM saves time by avoiding PCA calculation and operates on fewer numbers of nodes in the first two phases. Although the last training phase is basically identical for both algorithms and the learning length of the last training phase is much longer than the first two phases, GSOM still has 7%–15% of speed improvement for the overall training. It was reported that the use of this strategy involving SOM was used for analysing a large amount of eukaryote genomes and one of the highest performance supercomputers in the world was required [32]. For the large-scale analysis, which can take weeks to complete, 7%–15% of speed improvement means 1-2 days of time saving for a two-week computation.

Besides using Tables 6 and 7 to compare the training speed of SOM and GSOM, the tables also show that the training time grows exponentially from one order of oligonucleotide frequency to the higher-order oligonucleotide frequency. It is because of the rapid increment of dimensions when the order of oligonucleotide frequency increases.

## 4. DISCUSSION, CONCLUSION, AND FUTURE WORK

We have investigated four orders of oligonucleotide frequencies: di-, tri-, tetra-, and pentanucleotide frequencies on two artificial sets of species and a published simulated metagenomic dataset. Each of the two artificial sets contains 10 randomly selected species from NCBI database. We noticed that the F-measure can not distinguish the clustering quality for this application. Therefore, two methods have been defined for evaluating the performance of clustering DNA sequence fragments of species. This is done by observing the cluster formation with the labelled cluster map and then qualitatively and quantitatively comparing the LoM and IoM in the mixing region.

The results have shown that dinucleotide frequency is not a sufficiently strong signature for the tested 10 kb DNA sequences on the SOM-based algorithm. Similar to other reports, we also found that higher-order oligonucleotide frequencies, such as tri-, tetra-, and pentanucleotide frequencies, are carrying reasonably adequate genomic information to group intraspecies sequences and separate interspecies sequences [12, 19]; but the required computational power increases exponentially for each increased order of oligonucleotide frequency. Additionally, we noticed that increase of the order of oligonucleotide frequency may deteriorate the assignment of DNA sequence fragments to classes in some cases, which indicates the possible existence of optimal species-specific oligonucleotide frequency. For example, the trinucleotide frequency has a better discrimination power for *Acinetobacter* sp. ADP1 and *Bacillus subtilis* subsp. *subtilis* str. 168 then the tetra- and pentanucleotide frequencies. Therefore, analysts are recommended to start with trinucleotide frequency in large-scale projects and higher-order oligonucleotide frequencies may not always be better.

We also compare the SOM and GSOM algorithms for clustering the DNA sequence fragments of species. Both SOM and GSOM have shown similar results. However, in term of speed comparison, GSOM has more than 37% speed improvement over SOM in the first two training phases and

Table 6: Speed comparisons for the first two training phases of SOM and GSOM, in which the improvement columns represent the percentage of speed improvement for GSOM comparing to SOM.

| | Species Set 1 | | | Species Set 2 | | | simMC_Phrap | | |
|---|---|---|---|---|---|---|---|---|---|
| | SOM (sec) | GSOM (sec) | Improvement | SOM (sec) | GSOM (sec) | Improvement | SOM (sec) | GSOM (sec) | Improvement |
| Di | 54 | 34 | 37% | 24 | 15 | 38% | 2 | 1 | 50% |
| Tri | 188 | 115 | 39% | 74 | 45 | 39% | 7 | 4 | 43% |
| Tetra | 779 | 475 | 39% | 236 | 147 | 38% | 31 | 18 | 42% |
| Penta | 3031 | 1847 | 39% | 878 | 518 | 41% | 144 | 80 | 44% |

Table 7: Speed comparisons for the overall training time of SOM and GSOM, in which the improvement columns represent the percentage of speed improvement for GSOM comparing to SOM.

| | Species Set 1 | | | Species Set 2 | | | simMC_Phrap | | |
|---|---|---|---|---|---|---|---|---|---|
| | SOM (sec) | GSOM (sec) | Improvement | SOM (sec) | GSOM (sec) | Improvement | SOM (sec) | GSOM (sec) | Improvement |
| Di | 313 | 274 | 12% | 133 | 121 | 9% | 11 | 10 | 9% |
| Tri | 1048 | 942 | 10% | 427 | 387 | 9% | 39 | 36 | 8% |
| Tetra | 4639 | 3932 | 15% | 1297 | 1203 | 7% | 173 | 158 | 9% |
| Penta | 16839 | 15709 | 7% | 4702 | 4387 | 7% | 720 | 662 | 8% |

7%–15% speed improvement in the overall training. Therefore, GSOM is potentially a better alternative clustering tool. As a result of this study, we would suggest to use GSOM and a higher-order oligonucleotide frequency (at least trinucleotide frequency) to improve the strategy proposed by Abe et al. [15] for the binning process after WGS sequencing.

The method of combining oligonucleotide frequency and the SOM-based algorithm has provided a promising way of binning after WGS sequencing. However, there are limitations with this method. Since SOM-based algorithms are essentially data visualisation techniques, it is difficult to identify the exact cluster boundaries when clusters severely overlap with each other. The overlapping cluster can often be misinterpreted as a single cluster when no label is available. Therefore, a further development to overcome this cluster-overlapping problem is necessary for such SOM-based binning method to be fully practical. Additionally, at the current state, due to the high diversity of microbial communities and the nature of WGS sequencing, most of the unassembled sequences are less than 10 kb. In order to maximize the use of this binning strategy, more investigation on the optimal sequence length will need to be performed in the future work. On the other hand, the rapidly advancing sequencing technology and techniques that are capable of faster sequencing, higher coverage, and longer contig length are continuously being developed [33, 34]. The length of the unassembled fragment is expected to increase in the near future. Therefore, this binning strategy is useful for the analysis after WGS sequencing. Alternatively, when one is attempting to identify a specific species in the metagenome, which has already been sequenced, supervised learning methods can be applied. While PhyloPythia employs SVM as its supervised learning classifier, one can opt for other well-known supervised learning methods that has been used in other various applications [35, 36].

## REFERENCES

[1] G. W. Tyson, J. Chapman, P. Hugenholtz, et al., "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[2] J. C. Venter, K. Remington, J. F. Heidelberg, et al., "Environmental genome shotgun sequencing of the sargasso sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[3] S. G. Tringe, C. Von Mering, A. Kobayashi, et al., "Comparative metagenomics of microbial communities," *Science*, vol. 308, no. 5721, pp. 554–557, 2005.

[4] T. Woyke, H. Teeling, N. N. Ivanova, et al., "Symbiosis insights through metagenomic analysis of a microbial consortium," *Nature*, vol. 443, no. 7114, pp. 950–955, 2006.

[5] D. B. Rusch, A. L. Halpern, G. Sutton, et al., "The sorcerer *II* global ocean sampling expedition: northwest atlantic through eastern tropical pacific," *PLoS Biology*, vol. 5, no. 3, p. e77, 2007.

[6] S. Yooseph, G. Sutton, D. B. Rusch, et al., "The sorcerer *II* global ocean sampling expedition: expanding the universe of protein families," *PLoS Biology*, vol. 5, no. 3, p. e16, 2007.

[7] K. Chen and L. B. Pachter, "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLoS Computational Biology*, vol. 1, no. 2, p. e24, 2005.

[8] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," *PLoS Biology*, vol. 5, no. 3, p. e82, 2007.

[9] A. Rodriguez, Y. Zhang, N. Maltsev, and E. Marland, "Chisel—a framework for identification and characterization of taxonomic and phenotypic versions of enzymes," in *Proceedings of the Institute of Structural Molecular Biology (ISMB '06)*, Fortaleza, Brazil, 2006.

[10] N. Maltsev, M. Syed, A. Rodriguez, B. Gopalan, and F. Brockman, "A novel binning approach and its application to a metagenome from a multiple extreme environment," in *Proceedings of the Joint Genomics: GTL Awardee Workshop V and*

*Metabolic Engineering and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop*, North Bethesda, Md, USA, 2007.

[11] M. Huntemann, "MetaClust—entwicklung eines modularen Programms zum Clustern von Metagenomfragmenten anhand verschiedener intrinsischer DNA-Signaturen," Diploma thesis, University of Bremen, Germany, 2006.

[12] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[13] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 163, 2004.

[14] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," *Nature Methods*, vol. 4, no. 1, pp. 63–72, 2007.

[15] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, no. 4, pp. 693–702, 2003.

[16] S. Karlin, J. Mrázek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *Journal of Bacteriology*, vol. 179, no. 12, pp. 3899–3913, 1997.

[17] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity," *Current Opinion in Microbiology*, vol. 1, no. 5, pp. 598–610, 1998.

[18] C. Weinel, K. E. Nelson, and B. Tümmler, "Global features of the Pseudomonas putida KT2440 genome sequence," *Environmental Microbiology*, vol. 4, no. 12, pp. 809–818, 2002.

[19] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier," *Genome Research*, vol. 11, no. 8, pp. 1404–1409, 2001.

[20] Y. Z. Zhai, A. Hsu, and S. K. Halgamuge, "Scalable dynamic self-organising maps for mining massive textual data," in *Proceedings of the Lecture Notes in Computer Science (including subseries: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4234, LNCS III, pp. 260–267, Springer, Berlin, Germany, 2006.

[21] R. Amarasiri and D. Alahakoon, "Applying dynamic self organizing maps for identifying changes in data sequences," in *Design and Application of Hybrid Intelligent Systems*, pp. 682–691, IOS Press, Amsterdam, The Netherlands, 2003.

[22] S. Chen, D. Alahakoon, and M. Indrawan, "Background knowledge driven ontology discovery," in *Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, (EEE '05)*, pp. 202–207, 2005.

[23] H. Wang, F. Azuaje, and N. Black, "Improving biomolecular pattern discovery and visualization with hybrid self-adaptive networks," *IEEE Transactions on Nanobioscience*, vol. 1, no. 4, pp. 146–166, 2002.

[24] H. Wang, F. Azuaje, and N. Black, "Interactive GSOM-Based approaches for improving biomedical pattern discovery and visualization," in *Computational and Information Science*, vol. 3314 of *Lecture Notes in Computer Science*, pp. 556–561, Springer, Berlin, Germany, 2004.

[25] M. A. Karim, S. Halgamuge, A. J. R. Smith, and A. L. Hsu, "Manufacturing yield improvement by clustering," in *Neural Information Processing*, vol. 4234 of *Lecture Notes in Computer Science*, pp. 526–534, Springer, Berlin, Germany, 2006.

[26] A. L. Hsu, S.-L. Tang, and S. K. Halgamuge, "An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data," *Bioinformatics*, vol. 19, no. 16, pp. 2131–2140, 2003.

[27] A. L. Hsu and S. K. Halgamuge, "Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation," *International Journal of Approximate Reasoning*, vol. 32, no. 2-3, pp. 259–279, 2003.

[28] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 601–614, 2000.

[29] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 2nd edition, 1997.

[30] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, UK, 2nd edition, 1979.

[31] K. Mavromatis, N. Ivanova, K. Barry, et al., "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol. 4, no. 6, pp. 495–500, 2007.

[32] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, "Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes," *Gene*, vol. 365, no. 1-2, pp. 27–34, 2006.

[33] T. Jarvie, L. Du, and J. Knight, "Shotgun sequencing and assembly of microbial genomes: comparing 454 and Sanger methods," *Biochemica*, pp. 11–14, 2005.

[34] L. Bonetta, "Genome sequencing in the fast lane," *Nature Methods*, vol. 3, no. 2, pp. 141–146, 2006.

[35] S. K. Halgamuge and M. Glesner, "Fuzzy neural networks: between functional equivalence and applicability," *International Journal of Neural Systems*, vol. 6, no. 2, pp. 185–196, 1995.

[36] S. K. Halgamuge, "Self-evolving neural networks for rule-based data processing," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2766–2773, 1997.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 694297, 5 pages
doi:10.1155/2008/694297

*Research Article*

# Statistical Analysis of Twin Populations Using Dissimilarity Measurements in Hippocampus Shape Space

**Youngser Park,[1] Carey E. Priebe,[1] Michael I. Miller,[1] Nikhil R. Mohan,[1] and Kelly N. Botteron[2]**

[1] *Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218, USA*
[2] *Department of Psychiatry, Washington University, St. Louis, MO 63110, USA*

Correspondence should be addressed to Youngser Park, youngser@jhu.edu

Received 19 July 2007; Accepted 19 November 2007

Recommended by Daniel Howard

By analyzing interpoint comparisons, we obtain significant results describing the relationship in “hippocampus shape space” of clinically depressed, high-risk, and control populations. In particular, our analysis demonstrates that the high-risk population is closer in shape space to the control population than to the clinically depressed population.

## 1. INTRODUCTION

Major depressive disorder (MDD) is a mental disorder affecting about 16% of the US adult population, and is a major cause for concern not only in the United States but the world over. It is a disorder characterized by depressed mood, diminished interest or pleasure, significant weight loss, feelings of guilt or low self-worth, insomnia or hypersomnia, fatigue, poor concentration, or recurrent thoughts of death. The symptoms are widespread, and tend to be quite stable. In 2000, the World Health Organization (WHO) estimated depression to be the leading cause of disability as measured by years lived with disability (YLD) and the fourth leading contributor to the global burden of disease. See [1].

Over the years, a significant amount of research has been dedicated to finding physiological causes of MDD. One such study involved the catecholamine hypothesis [2] that suggested that MDD is caused by decreased levels of the neurotransmitters norepinephrine and serotonin. This finding led to most modern day medication for MDD, which works by preventing the reuptake of these neurotransmitters. Neuroimaging research has also shown that enlarged ventricles, sulci, reduced volume of the frontal lobe and basal ganglia are also associated with depressive episodes [2].

The studies aforementioned involved studying the brain once MDD had already set in. The physiological changes are associated with the symptoms themselves. What about physiological *predictors* for MDD? Such predictors would facilitate the diagnosis of the disorder well before the onset of the symptoms, perhaps allowing measures to prevent the symptoms from ever appearing.

A vast amount of research is being conducted in order to find biological predispositions to MDD. There is evidence correlating shape differences of the hippocampus to depression [3] and schizophrenia [4]. In this manuscript we analyze interpoint comparisons [5] to investigate the relationship in “hippocampus shape space” of three populations among twins. The subjects are categorized into three categories: the affected subjects (clinically depressed, or MDD), the nonaffected cotwin of the MDD subjects (high-risk, or HR), and the nonaffected twin pair (Control, or CTRL). The dataset includes both monozygotic (MZ) and dizygotic (DZ) twin subjects.

According to established literature, the concordance rate for monozygotic (MZ) HR subjects is 40%, and for dizygotic (DZ) HR subjects 11% [6]. This demonstrates that the subjects labeled HR (due to the fact that their twin is MDD) are in fact high risk—they develop MDD at a higher rate than the general population.

## 2. DATA

Our data set includes $n = 114$ subjects (57 twin pairs): 29 CTRL-CTRL pairs, 22 HR-MDD pairs, and 6 MDD-MDD pairs. The subjects are young female twins recruited through an epidemiological sample based on Missouri birth records.

To ensure that hippocampus shape space is the only independent variable, other factors had to be controlled; all of the subjects were right handed and were screened for factors that may cause structural changes of the brain such as loss of consciousness greater than 5 minutes, chronic medical or neurological illnesses, or pregnancy.

To obtain images of the hippocampus, very high resolution magnetic resonance imaging (MRI) scans were required. The Siemens Vision/Sonata 1.5T scanner was used to acquire three MPRAGE scans [7] (160 slices at 256*x*256 FoV, 1.0 mm$^3$ isotropic voxels). Using Analyze [8], the images were registered and averaged, converted to 8-bits while optimizing the intensity range, and interpolated to 0.5 mm isotropic voxels. The image protocol implemented above allows for optimal comparative analysis.

For each of left and right hemispheres separately, 22 three-dimensional landmarks were identified for each hippocampus and were used to generate and align hippocampal subcubes to a standardized orientation. It is these landmark data that we employ herein.

## 3. SHAPE

Using the landmark data, for each pair of subjects and for each of left and right hippocampus, we produce an interpoint shape comparison, as described below.

For two subjects $x$ and $y$ (for the left hemisphere, say), let $x_1, \dots, x_N$ and $y_1, \dots, y_N$ be the corresponding landmarks, where $N = 22$.

### 3.1. Landmark matching

Finding the shape comparison involves a landmark matching (LM) transformation. The transformation is nonparametric, and this flexibility implies that overfitting must be guarded against via regularization. LM finds a diffeomorphism $\varphi$ that minimizes an error criterion which includes both landmark mismatch and transformation complexity. That is,

$$\varphi^* = \arg\inf_{\varphi} \sigma\, d(\varphi_{\text{identity}}, \varphi)^2 + \sum_{i=1}^{N} \|\varphi(x_i) - y_i\|^2, \qquad (1)$$

where $d$ is a geodesic distance in a group of diffeomorphisms [9] and $\sigma > 0$ is a regularization parameter which controls the relative contribution of transformation complexity versus landmark mismatch to the optimization objective. The algorithm solves the nonlinear Euler equation by a Newton method combined with a shooting procedure [10].

We use $\text{LM}(x, y) = \|\varphi^*\|$, the energy of the minimizing diffeomorphism, as the shape comparison between two subjects $x$ and $y$ (for the left hemisphere, say).

### 3.2. Interpoint comparison matrices

Applying LM to the left or right hippocampus data for each pair of subjects yields an interpoint comparison matrix $\widetilde{D}$. However, $\widetilde{D}$ is $n \times n$, hollow (zeros on the diagonal) and is asymmetric. That is, we obtain matrices $\widetilde{D}_{\text{LM-Left}}$ and $\widetilde{D}_{\text{LM-Right}}$.

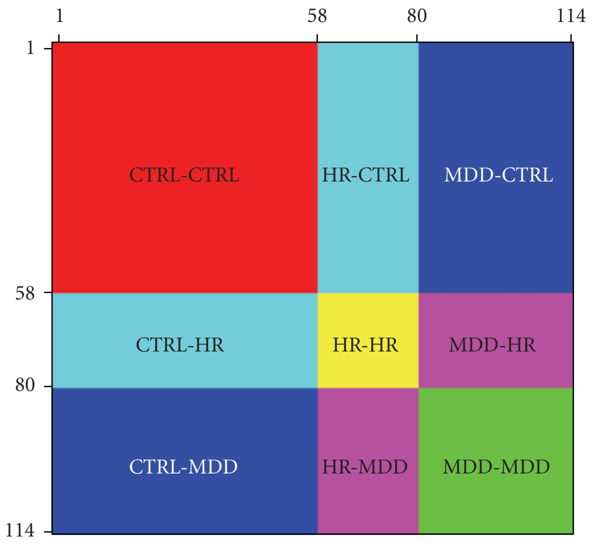


FIGURE 1: Structure of the interpoint comparison matrices $D$ for the 114 subjects.

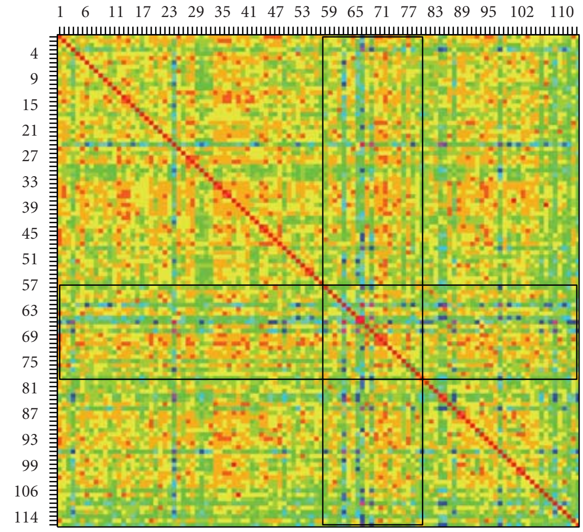


FIGURE 2: The interpoint comparison matrix $D_{\text{LM-Left}}$, after symmetrization, for the 114 subjects. The comparison values are color-coded, with red representing zero (e.g., the diagonal entries) to green representing large values.

The nature of the hippocampus shape space is such that under ideal conditions, it should yield a symmetric distance matrix. The asymmetry of the matrix $\widetilde{D}$ does not reflect the true nature of the hippocampus shape space, and is in fact a result of the limitations in the LM matching method. Hence, before further investigation, $\widetilde{D}$ must be symmetrized to $D$, using an appropriate symmetrization technique [11]. In this work we symmetrize via $d_{ij} = \min\{\tilde{d}_{ij}, \tilde{d}_{ji}\}$.

Figure 1 depicts the structure of the interpoint comparison matrices for the 114 subjects. Figure 2 depicts the actual interpoint comparison matrix $D_{\text{LM-Left}}$ (after symmetrization) for the 114 subjects.

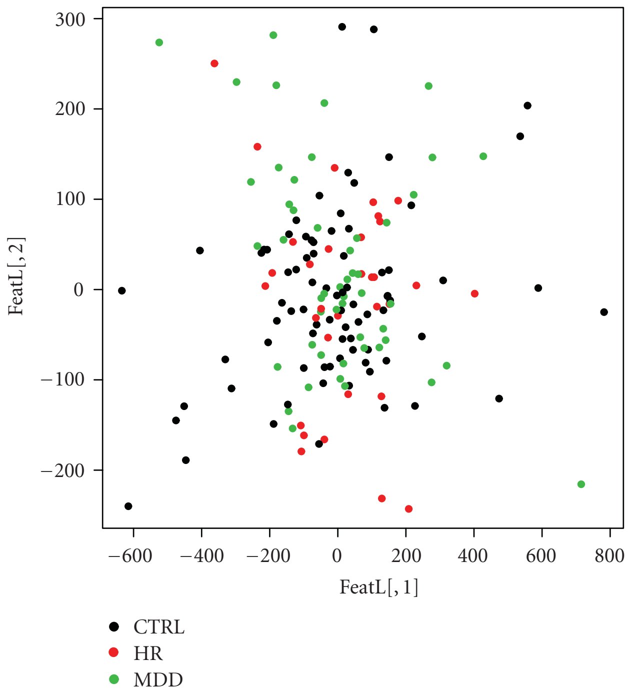


FIGURE 3: A multidimensional scaling scatter plot of $D_{\text{LM-Left}}$ mapped into $\mathbb{R}^2$. Little can be discerned from this plot regarding the relationship in hippocampus shape space of the three populations (MDD, HR, CTRL)—no class-conditional differentiation is apparent.

## 4. STATISTICAL ANALYSIS

Our task is to begin describing the relationship of the three populations (MDD, HR, CTRL) amongst one another in the hippocampus shape space elicited by the LM interpoint comparisons. First, we present a multidimensional scaling (MDS) [12] scatter plot; unfortunately, we see in Figure 3 that no significant relationship can be discerned from this plot. Employing linear discriminant analysis (LDA) after MDS for all possible MDS target dimensionalities—analysis via LDA $\circ$ MDS $\circ$ LM $(\cdot)$ *a la* Miller et al. [13]—yields no classification capabilities statistically significantly superior to chance. Nevertheless, we will see in Figures 4 and 5 a suggestion that perhaps progress can be made on our task, given a sufficiently clever methodology.

Figure 4 depicts kernel probability density estimates [14] for the LM-Left comparisons to show that the entries of the interpoint comparisons matrix $D_{\text{LM-Left}}$ that correspond to comparisons between HR and CTRL (the solid line in Figure 4) are, overall, *smaller* than the entries which correspond to comparisons between HR and MDD. That is, Figure 4 suggests a *stochastic ordering* relationship [15]: $d(\text{HR}, \text{CTRL}) <^{\text{st}} d(\text{HR}, \text{MDD})$. Such a result is precisely what we seek. Again, dependencies amongst the entries of $D$ make it difficult to assess the statistical significance of the result depicted in Figure 4.

Each row of the interpoint comparisons matrix $D$, corresponding to a single HR subject, gives rise to two samples: $\{d(\text{HR}, \text{CTRL}_i)\}$ and $\{d(\text{HR}, \text{MDD}_j)\}$. That is, we have the vector of comparisons from that HR subject to every CTRL subject, and we have the vector of comparisons from that HR subject to every MDD subject. (We do not include in these vectors the twin of the particular HR subject under consideration; ignoring twinnedness in the analysis proves beneficial that we eliminate bias in similarity status between a subject and her twin that is not due to

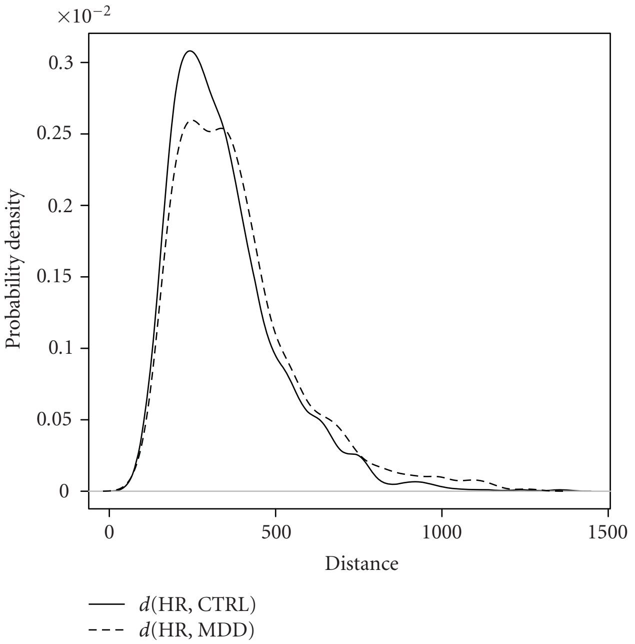


FIGURE 4: This figure shows kernel probability density estimates for $D_{\text{LM-Left}}$. The solid line depicts $d(\text{HR}, \text{CTRL})$ and the dashed line depicts $d(\text{HR}, \text{MDD})$.

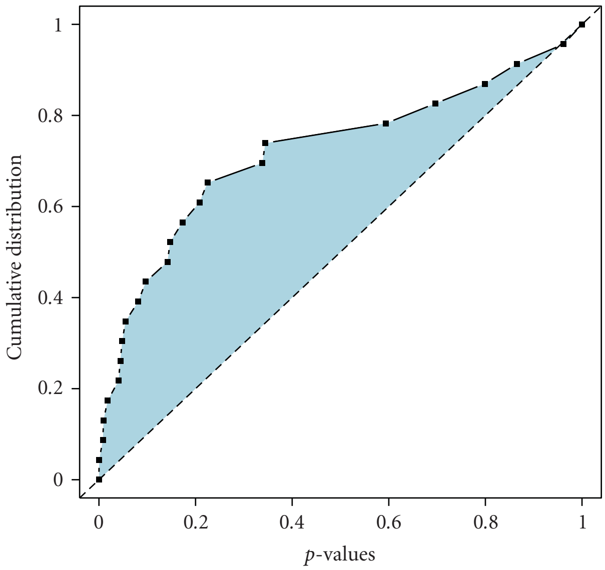


FIGURE 5: This figure shows the quantile-quantile plot for $D_{\text{LM-Left}}$. Depicted are the individual $P$-values for a Wilcoxon-Mann-Whitney test of each HR subject, in turn, based on the two samples $\{d(\text{HR}, \text{CTRL}_i)\}$ and $\{d(\text{HR}, \text{MDD}_i)\}$.

condition (MDD,HR,CTRL).) For this individual HR subject's two sample data, a Wilcoxon-Mann-Whitney test [16] of the null hypothesis that the distribution of comparisons $d(\text{HR}, \text{CTRL})$ is the same as the distribution of comparisons $d(\text{HR}, \text{MDD})$, against the alternative of *stochastic ordering*, yields a $P$-value. Figure 5 provides a quantile-quantile plot of these $P$-values for $D_{\text{LM-Left}}$. Under the null hypothesis, these $P$-values would be expected to be distributed approximately uniform(0,1). The plot demonstrates a clear deviation from a uniform distribution, again suggesting a *stochastic ordering* relationship—$d(\text{HR}, \text{CTRL}) <^{\text{st}} d(\text{HR}, \text{MDD})$. Again, dependencies amongst the entries of $D$ make it difficult to assess the statistical significance of the result depicted in Figure 5.

The quantile-quantile plot independently reiterates the suggestion of a stochastic ordering relationship that was first seen using the kernel probability density estimates. Thus, while Figures 4 and 5 give an inkling of the type of information that can be gleaned regarding the relationship in hippocampus shape space of the three populations (MDD, HR, CTRL) amongst one another, it remains henceforth to accurately assess the Figures' suggestion.

## 5. CLASSIFICATION

To further uncover the characteristics of hippocampus shape space, we consider the task of *classifying* each HR subject as either MDD or CTRL.

As before, we consider the two samples, $\{d(\text{HR}, \text{CTRL}_i)\}$ and $\{d(\text{HR}, \text{MDD}_j)\}$, associated with each individual HR subject. We classify the HR subject as belonging to MDD or CTRL based on the Wilcoxon-Mann-Whitney test statistic $P$-value, as described in [17]; (see also [15, page 183]).

Once we have classified each of the HR subjects in this way, we assess the relative similarity of HR to CTRL versus MDD based on the classifier's performance—based on the collection of HR subjects' classifier-assigned class labels, taken as a whole.

This procedure can be employed with LM interpoint comparisons obtained on Left, Right, or both Left and Right hippocampuses, and with any of the three populations (HR, CTRL, MDD) as the population of interest—the role of HR in the description above.

## 6. RESULTS

Classifying the 22 HR subjects as either MDD or CTRL using $D_{\text{LM-Left}}$ results in 19 classified as CTRL versus 3 classified as MDD. The probability of obtaining a result this extreme or more extreme (the $P$-value) under the least favorable null hypothesis $H_0$ : *HR are equally likely to be classified as MDD as CTRL* is $P < .01$ against each one-sided alternative. LM-Right yields 16 classified as CTRL versus 6 classified as MDD—classification performance not statistically significantly distinguishable from chance. Combining left and right, the shape comparisons LM (LM-Left and Right) yields 20 classified as CTRL versus 2 classified as MDD—$P < .0005$ for each one-sided alternative and strong statistical evidence that HR is more like CTRL than MDD in hippocampus shape space.

Figure 6: Artist's rendition of what hippocampus shape space might look like were it one-dimensional—if the population shapes could be accurately represented in $\mathbb{R}^1$. Our results suggest that the joint relationship of the three populations, in terms of shape, puts the CTRL population *between* the HR and MDD populations. The relationship depicted here holds for both LM-Left and LM-Right, although our results suggest that for the left hippocampus CTRL is shifted closer to HR while for the right hippocampus the CTRL is shifted closer to MDD.

An analogous analysis—classifying the 33 MDD subjects as either HR or CTRL using LM-Left—shows that MDD is more like CTRL than it is like HR ($P < .00002$ for each one-sided alternative), and that the left carries more information than does the right—the $P$-values are smaller indicating that the signal is stronger.

The results obtained from classifying the 59 CTRL subjects as either HR or MDD are more nuanced: in this case, using LM-Left indicates that CTRL is more like HR than it is like MDD ($P < .0005$) while using LM-Right indicates that CTRL is more like MDD than it is like HR ($P < .0005$). This hemispherical ambiguity provides further insight into hippocampus shape space.

Finally, we note that in the last two columns of Table 1 we consider classifying the 22 HR subjects (via leave-one-out crossvalidation) as HR or MDD and as HR or CTRL. These results are consistent with our other findings—HR is more difficult to distinguish from CTRL than from MDD, and the information extracted via LM-Left is more powerful for this task than is LM-Right.

## 7. CONCLUSIONS/DISCUSSION

Our analysis indicates that HR is more like CTRL than it is like MDD, MDD is more like CTRL than it is like HR, and CTRL is not obviously more like one or the other. Also, we discern that the left hippocampus carries more information than does the right.

If hippocampus shape space were one-dimensional—if the population shapes could be accurately represented in $\mathbb{R}^1$—then the joint relationship described by these three results could be depicted as in Figure 6, with the CTRL population *between* the HR and MDD populations in terms of shape. However, it must be noted that this depiction (Figure 6) offers only a simplified view of the true infinite dimensional nature of the shape space configuration, as suggested by the fact that the 2-dimensional MDS embedding depicted in Figure 3 presents little or no class separation.

### 7.1. On Populations

Our stated task is in terms of *populations*—to begin describing the relationship in hippocampus shape space of the three populations (MDD, HR, CTRL) amongst one another. However, our results are *conditional*—using LM-Left we classify, for example, the 22 HR subjects representing the HR

Table 1: Output of classifier based on the Wilcoxon-Mann-Whitney test statistic. For example, the first numerical column, "H : CvM," gives the number of HR classified as CTRL versus MDD and the second numerical column, "H : MvC," gives the number of HR classified as MDD versus CTRL. Thus, we find that combining left and right, the shape comparisons LM (LM-Left and Right) yields 20 HR classified as CTRL versus 2 HR classified as MDD—strong statistical evidence that HR is more like CTRL than MDD in hippocampus shape space. (This analysis is based on 22 HR subjects, 33 MDD subjects, and 59 CTRL subjects. Thus, e.g., the two HR numbers, H : CvM and H : MvC, should sum to 22. Discrepancies are due to situations in which the classifier makes "no decision" as described in [17]; (see also [15, page 183]).

| | H : CvM | H : MvC | M : HvC | M : CvH | C : HvM | C : MvH | H : HvM | H : HvC |
|---|---|---|---|---|---|---|---|---|
| LM-Left | 19 | 3 | 5 | 28 | 48 | 11 | 16 | 7 |
| LM-Right | 16 | 6 | 9 | 24 | 22 | 33 | 13 | 8 |
| **LM-Left** & **Right** | 20 | 2 | 6 | 27 | 31 | 25 | 16 | 6 |

population as belonging to either the MDD or the CTRL class, conditionally on "training" data from MDD and CTRL. This, in fact, is the standard approach in probabilistic pattern recognition; see, for example, [18]. The difference between a focus on populations versus conditionals is indicative of a difference between "policy science" and "laboratory science" [19]. A justification for the conditional approach in "laboratory science" is given in [18] where it is claimed that the unconditional approach " ... would be unnatural, because in a given application, one has to live with the [training] data at hand." In "policy science", however, knowledge about the populations themselves may be the focus.

By performing our analysis thrice, for each of the three populations in turn conditionally on the "training" data from the other two, we obtain three conditionals. Letting $n_j$ denote the class-conditional sample sizes for each of the three classes, we see that the joint distribution for our sample is $(n_1+n_2+n_3)\cdot d$-dimensional (where $d$ is the presumed "shape space" dimensionality of each observation). Each conditional considered is $n_j\cdot d$-dimensional, with one population remaining. The overall joint distribution of interest—the three populations in "shape space"—is of course not simply the product of our three conditionals. However, *some* population inferences regarding stochastic ordering can be performed via the (multiple) conditionals, and in particular the conditional approach justifies the simplistic view of our three populations in "shape-space" given by Figure 6.

## ACKNOWLEDGMENT

Sincere appreciation to Michael Bowers (JHU), Timothy Brown (JHU), Anthony Kolasny (JHU), Tomoyuki Nishino (WashU), and the others for their valuable assistance.

## REFERENCES

[1] "Depression. National Institute of Mental Health," http://www.nimh.nih.gov/publicat/depression.cfm.

[2] L. B. Alloy, J. H. Riskind, and M. J. Manos, *Abnormal Psychology: Current Perspectives*, McGraw-Hill, New York, NY, USA, 9th edition, 2005.

[3] J. Posener, L. Wang, J. Price, et al., "High dimensional mapping of the hippocampus in depression," *American Journal of Psychiatry*, vol. 160, no. 1, pp. 83–89, 2003.

[4] J. G. Csernansky, S. Joshi, L. Wang, J. W. Haller, M. Gado, J. P. Miller, U. Grenander, and M. I. Miller, "Hippocampal morphometry in schizophrenia by high dimensional brain mapping," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 19, pp. 11406–11411, 1998.

[5] J.-F. Maa, D. K. Pearl, and R. Bartoszyński, "Reducing multidimensional two-sample data to one-dimensional interpoint comparisons," *Annals of Statistics*, vol. 24, no. 3, pp. 1069–1074, 1996.

[6] "Genetics and major psychiatric disorders:a program for genetic counselors. National Coalition for Health," http://www.nchpeg.org/cdrom/empiric.html.

[7] J. P. Mugler III and J. R. Brookeman, "Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE)," *Magnetic Resonance in Medicine*, vol. 15, no. 1, pp. 152–157, 1990.

[8] Analyze Software, "Mayo Clinic," http://www.mayo.edu/bir/Software/Analyze/.

[9] M. I. Miller, A. Trouve, and L. Younes, "Geodesic shooting for computational anatomy," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 2, pp. 209–228, 2006.

[10] S. Allassonnire, A. Trouve, and L. Younes, "Geodesic shooting and diffeomorphic matching via textures meshes," in *Proceedings of the 5th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR '05)*, pp. 365–381, Augustine, Fla, USA, November 2005.

[11] T. Saito and H. Yadohisa, *Data Analysis of Asymmetric Structures*, Marcel Dekker, New York, NY, USA, 2005.

[12] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, New York, NY, USA, 2nd edition, 2001.

[13] M. Miller, C. Priebe, and Y. Park, "Collaborative computational anatomy: the perfect storm for mri morphometry study of the human brain via diffeomophic metric mapping, multidimensional scaling and linear discriminant analysis," to appear in *Proceedings of the National Academy of Science*.

[14] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York, NY, USA, 1986.

[15] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1, Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2005.

[16] J. Rice, *Mathematical Statistics and Data Analysis*, Addison-Wesley, Reading, Mass, USA, 2nd edition, 1995.

[17] C. E. Priebe, "Olfactory classification via interpoint distance analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 404–413, 2001.

[18] L. Devroye, L. Gyorfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Number 31 in Applications of mathematics, 1996.

[19] B. Caffo, Personal communication, 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 784354, 5 pages
doi:10.1155/2008/784354

*Research Article*

# An Approximate Numerical Technique for Characterizing Optical Pulse Propagation in Inhomogeneous Biological Tissue

**Chintha C. Handapangoda and Malin Premaratne**

*Advanced Computing and Simulation Laboratory (AXL), Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia*

Correspondence should be addressed to Chintha C. Handapangoda, chintha.handapangoda@eng.monash.edu.au

Received 8 September 2007; Accepted 19 December 2007

Recommended by Daniel Howard

An approximate numerical technique for modeling optical pulse propagation through weakly scattering biological tissue is developed by solving the photon transport equation in biological tissue that includes varying refractive index and varying scattering/absorption coefficients. The proposed technique involves first tracing the ray paths defined by the refractive index profile of the medium by solving the eikonal equation using a Runge-Kutta integration algorithm. The photon transport equation is solved only along these ray paths, minimizing the overall computational burden of the resulting algorithm. The main advantage of the current algorithm is that it enables to discretise the pulse propagation space adaptively by taking optical depth into account. Therefore, computational efficiency can be increased without compromising the accuracy of the algorithm.

## 1. INTRODUCTION

Modeling of light propagation through biological tissues is important for many medical applications such as optical tomography for cancer detection [1] and noninvasive detection of diabetes mellitus [2]. Researchers have been working on modeling biological tissues over the last two decades [3, 4].

Light propagation through biological tissues can be modeled using the photon transport equation (PTE) [5, 6]. Several numerical models have been developed to solve the PTE over the recent years [7–10]. These models include techniques for solving the steady-state PTE [7], as well as the transient PTE for short pulse propagation [8–10]. Several different variations of PTE for inhomogeneous media have been proposed in the literature [6, 11–15]. However, most of these variations are not a result of fundamental errors or differences but due to different assumptions about the medium or wave field properties. For example, [6, 11, 13] look at spatially slowly varying isotropic refractive index profiles in their work. Interestingly, the approach given in [15] is formulated to accommodate geometric optics approximations but ignore the wavefront curvature in their derivation. Wavefront curvature in the context of slowly varying refractive index approximation is considered in [6]. Numerical considerations necessary to account for such slowly varying spatial refractive profiles are considered in [12, 14].

This paper presents a technique for modeling the light propagation through weakly scattering and absorbing media by solving the three-dimensional transient PTE numerically. In a weakly scattering medium, the ray paths can be approximated by the Eikonal equation [16]. First, a set of ray paths is calculated depending on the refractive index profile of the medium. There are several existing methods to do this [16–18]. In this paper, we have briefly described a ray tracing technique proposed by Sharma et al. [16]. The next step is to solve the RTE written in terms of the arc length [6] on each of these paths. The Laguerre Runge-Kutta-Fehlberg method [19] can be used for this purpose.

This paper is organized in four sections. Section 2 presents the formulation of the numerical technique. Section 3 contains some results obtained using this technique and a discussion on these results. Section 4 includes the conclusions.

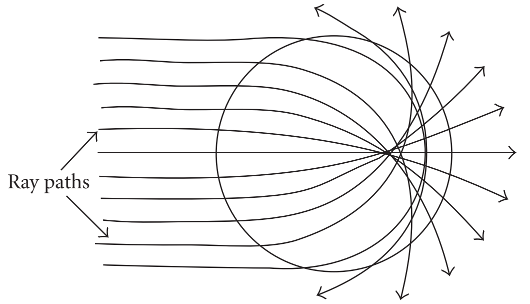


FIGURE 1: A set of possible ray paths in a Maxwell's fish-eye.

## 2. FORMULATION

The photon transport equation for a medium with a spatially varying isotropic refractive index, in standard spherical coordinates, is [6]

$$\begin{aligned}&\frac{n(\mathbf{r})}{c}\frac{\partial}{\partial t}I(\mathbf{r},\mathbf{\Omega},t)+\left(\frac{1}{R_1(s)}+\frac{1}{R_2(s)}\right)I(\mathbf{r},\mathbf{\Omega},t)\\&\quad+n^2(\mathbf{r})\frac{\partial}{\partial s}\left(\frac{I(\mathbf{r},\mathbf{\Omega},t)}{n^2(\mathbf{r})}\right)+\sigma_t(\mathbf{r})I(\mathbf{r},\mathbf{\Omega},t)\\&=\sigma_s(\mathbf{r})\int_{4\pi}P(\mathbf{\Omega},\mathbf{\Omega}')I(\mathbf{r},\mathbf{\Omega}',t)d\mathbf{\Omega}'+F(\mathbf{r},\mathbf{\Omega},t),\end{aligned}\tag{1}$$

where $\mathbf{r}$ is the position vector of a point on a path of a ray, $s$ is the arc length along a ray, $\mathbf{\Omega}=d\mathbf{r}/ds$, $t$ is the time variable, $I(\mathbf{r},\mathbf{\Omega},t)$ is the intensity, $n(\mathbf{r})$ is the refractive index profile, $c$ is the speed of light in vacuum, $R_1(s)$ and $R_2(s)$ are the principal radii of curvatures of the geometrical wave-fronts, $\sigma_t$ is the attenuation coefficient with $\sigma_t=\sigma_a+\sigma_s$, $\sigma_a$ is the absorption coefficient and $\sigma_s$ is the scattering coefficient, $P(\mathbf{\Omega},\mathbf{\Omega}')$ is the phase function, and $F(\mathbf{r},\mathbf{\Omega},t)$ represents sources inside the medium.

The path of rays in a medium with a spatially varying refractive index is given by the Eikonal equation [16–18]:

$$\frac{d}{ds}\left(n(\mathbf{r})\frac{d\mathbf{r}}{ds}\right)=\nabla n(\mathbf{r}),\tag{2}$$

where $\mathbf{r}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$, $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors along $x$, $y$, and $z$ axes, respectively, and $\nabla=((\partial/\partial x)\mathbf{i}+(\partial/\partial y)\mathbf{j}+(\partial/\partial z)\mathbf{k})$. We use the ray model of light here, based on geometric optic techniques, which neglects the concept of wavelength $\lambda$ (i.e., $\lambda\to 0$) [20]. Thus, this approximation will be valid for modeling light propagation through inhomogeneous media in which the properties vary very slowly compared to the wavelength of the incident light. Also, the geometric optic techniques assume that the field behaves locally as a plane wave and that the intensity does not change rapidly [20].

The technique proposed in this paper is to first solve (2) to obtain a set of possible ray paths and then solve the PTE, (1), for each of these paths. The main advantage of the current algorithm is that it enables to discretise the pulse propagation space adaptively by taking optical depth into account. Therefore, computational efficiency can be increased without compromising the accuracy of the algorithm [21]. Several techniques had been introduced for solving the Eikonal equation by various research groups [16–18]. We adopt the technique proposed by Sharma et al. [16] because it uses Runge-Kutta integration, which will be used again later to solve photon transport equation in discrete ordinate method setting. In this method, a new variable $q$ is introduced such that $dq=ds/n$. Then, (2) can be written as

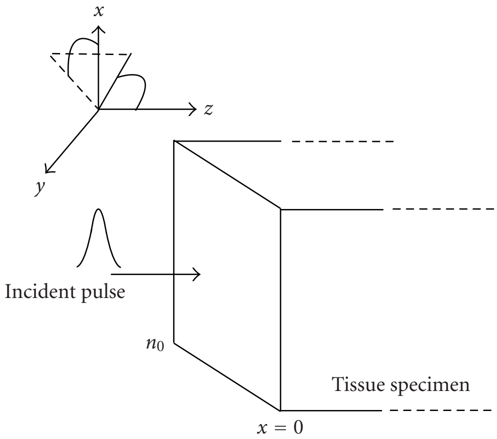


FIGURE 2: Light pulse incident on a tissue specimen.

$$\frac{d^2\mathbf{r}}{dq^2}=n(\mathbf{r})\nabla n(\mathbf{r}).\tag{3}$$

The optical ray vector is defined as

$$\mathbf{Q}=\frac{d\mathbf{r}}{dq}=n\xi\mathbf{i}+n\eta\mathbf{j}+n\mu\mathbf{k},\tag{4}$$

where $\xi=\sin\theta\cos\phi$, $\eta=\sin\theta\sin\phi$, and $\mu=\cos\theta$. Equation (3) can be written in matrix format as

$$\frac{d^2R}{dq^2}=D(R),\tag{5}$$

where

$$R=\begin{pmatrix}x\\y\\z\end{pmatrix},\qquad Q=\begin{pmatrix}n\xi\\n\eta\\n\mu\end{pmatrix},\qquad D=n\begin{pmatrix}\dfrac{\partial n}{\partial x}\\\dfrac{\partial n}{\partial y}\\\dfrac{\partial n}{\partial z}\end{pmatrix}.\tag{6}$$

Thus, (5) can be solved using the Runge-Kutta algorithm starting from a known point $(R_0,Q_0)$. That is, (5) with the initial condition $(R_0,Q_0)$ will successively generate points $(R_1,Q_1),(R_2,Q_2),\ldots,(R_n,Q_n)$ along the path [16]. Therefore, if we select a set of starting points and initial directions, $Q_0$, we can obtain a set of ray paths by numerically integrating (5). For example, such ray tracing for Maxwell's fish-eye gives the paths as shown in Figure 1 [22].

Next step is to solve the PTE, (1), for a weakly scattering medium on each of the above sets of paths, by numerically integrating (5). While tracing the ray paths from the above algorithm, the PTE can be solved to find the intensity at each point on the path.

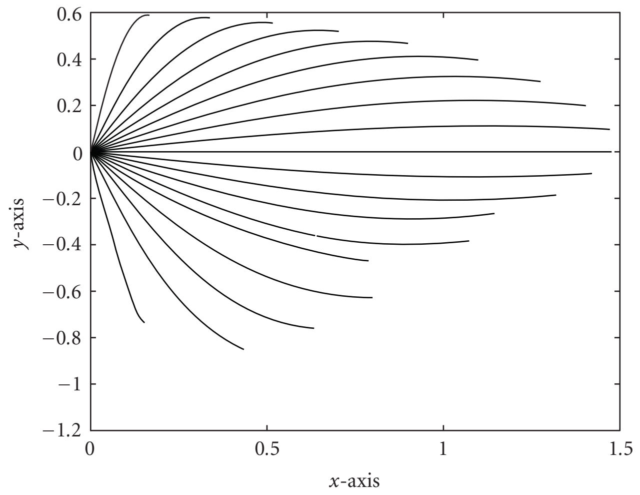


FIGURE 3: Few ray paths for a medium with a refractive index profile given by (14).

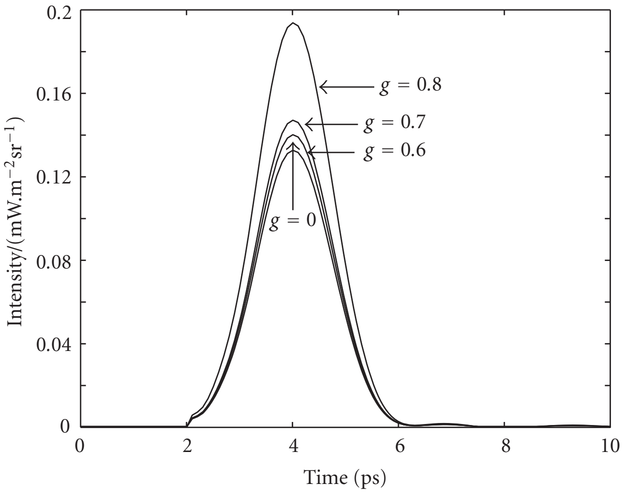


FIGURE 4: Intensity a $z = 1$ mm for different asymmetry factor ($g$) values.

First, we use the following transformation which maps the PTE to a moving coordinate system on ray paths:

$$\tau = t - \frac{s}{v}, \tag{7}$$

where $v$ is the speed of light inside the medium along the ray path. Using (7) in (1) results in

$$\begin{aligned} n^2(\mathbf{r})\frac{\partial}{\partial s}\left(\frac{I(\mathbf{r},\mathbf{\Omega},\tau)}{n^2(\mathbf{r})}\right) + \left(\frac{1}{R_1(s)}+\frac{1}{R_2(s)}\right)I(\mathbf{r},\mathbf{\Omega},\tau) \\ + \sigma_t(\mathbf{r})I(\mathbf{r},\mathbf{\Omega},\tau) \\ = \sigma_s(\mathbf{r})\int_{4\pi} P(\mathbf{\Omega},\mathbf{\Omega}')I(\mathbf{r},\mathbf{\Omega}',\tau)d\mathbf{\Omega}' + F(\mathbf{r},\mathbf{\Omega},\tau). \end{aligned} \tag{8}$$

In this paper, we consider plane waves so that the second term in the left-hand side of (8) vanishes. That is,

$$\begin{aligned} n^2(\mathbf{r})\frac{\partial}{\partial s}\left(\frac{I(\mathbf{r},\mathbf{\Omega},\tau)}{n^2(\mathbf{r})}\right) + \sigma_t(\mathbf{r})I(\mathbf{r},\mathbf{\Omega},\tau) \\ = \sigma_s(\mathbf{r})\int_{4\pi} P(\mathbf{\Omega},\mathbf{\Omega}')I(\mathbf{r},\mathbf{\Omega}',\tau)d\mathbf{\Omega}' + F(\mathbf{r},\mathbf{\Omega},\tau). \end{aligned} \tag{9}$$

The Laguerre Runge-Kutta-Fehlberg (LRKF)method [19] can be used to solve (9) for the intensity at selected points on each ray path. The LRKF method is briefly described below. Since we solve (9) on a known ray path at a known point $n(\mathbf{r})$, $\sigma_t(\mathbf{r})$, and $\sigma_s(\mathbf{r})$ are constants at that point. First, we use Gaussian quadrature [23] to approximate the integral:

$$\begin{aligned} n^2(\mathbf{r})\frac{\partial}{\partial s}\left(\frac{I(\mathbf{r},\mathbf{\Omega},\tau)}{n^2(\mathbf{r})}\right) + \sigma_t(\mathbf{r})I(\mathbf{r},\mathbf{\Omega},\tau) \\ = \sigma_s(\mathbf{r})\sum_{j=1}^{q} w_j P(\mathbf{\Omega},\mathbf{\Omega}_j)I(\mathbf{r},\mathbf{\Omega}_j,\tau) + F(\mathbf{r},\mathbf{\Omega},\tau), \end{aligned} \tag{10}$$

where $\mathbf{\Omega}_j$ is the $j$th quadrature point and $w_j$ is the corresponding Gaussian weight [23]. Then, the time dependence is represented using a Laguerre expansion [24]:

$$I(\mathbf{r},\mathbf{\Omega},\tau) = \sum_{k=0}^{N} B_k(\mathbf{r},\mathbf{\Omega})L_k(\tau), \tag{11}$$

where

$$\int_0^\infty L_n(\tau)L_m(\tau)e^{-\tau}d\tau = \delta_{mn}. \tag{12}$$

Using (11) in (10) and taking moments, we get

$$\begin{aligned} n^2(\mathbf{r})\frac{\partial}{\partial s}\left(\frac{B_n(\mathbf{r},\mathbf{\Omega})}{n^2(\mathbf{r})}\right) + \sigma_t(\mathbf{r})B_n(\mathbf{r},\mathbf{\Omega}) \\ = \sigma_s(\mathbf{r})\sum_{j=1}^{q} w_j P(\mathbf{\Omega},\mathbf{\Omega}_j)B_n(\mathbf{r},\mathbf{\Omega}_j) + F_n(\mathbf{r},\mathbf{\Omega}), \end{aligned} \tag{13}$$

where $F_n(\mathbf{r},\mathbf{\Omega})$ is the Laguerre coefficient of the source term $F(\mathbf{r},\mathbf{\Omega},\tau)$. Since we have previously traced the ray paths and chosen a set of points $\mathbf{r}$ and $\mathbf{\Omega}$ values are known, thus, (13) can be solved using the Runge-Kutta-Fehlberg algorithm [25].

## 3. RESULTS AND DISCUSSION

Figure 3 shows few ray paths for a medium with a refractive index profile given by

$$n = n_0 e^{-x^2}, \tag{14}$$

where $n_0 = 2$. Figures 4 and 5 were obtained from the above algorithm. We have obtained these results for a pulse propagating through a single ray path. The Henyey-Greenstein phase function [27] was used for the simulation where $g$ is the asymmetry factor.

Figure 4 shows the variation of intensity with time at $z = 1$ mm with varying $g$. The graphs correspond to $g = 0.8$, $g = 0.7$, $g = 0.6$, and $g = 0$. Other parameters such as the scattering coefficient and the absorption coefficient were kept constant for all the three graphs. The condition $g = 0$ corresponds to the isotropic scattering case while $g = 0.8$ represents forward scattering. This is illustrated by the above four graphs.

Figure 5 shows the variation of the forward intensity at different locations, that is, corresponding to different $z$ values

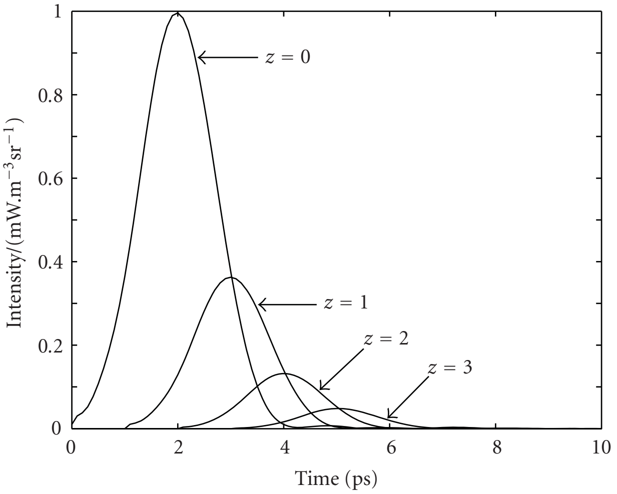


Figure 5: Forward intensity at different locations for isotropic ($g = 0$) scattering.

(in mm), with a constant asymmetry factor $g = 0$. It can be clearly seen from this figure that the intensity reduces with increasing distance due to scattering and absorption. Also, the pulse is shifted in time as shown.

## 4. CONCLUSION

This paper introduced a novel approximate numerical model to solve the photon transport equation with inhomogeneous refractive index and inhomogeneous scattering and absorption cross-sections for weakly scattering biological tissue. The proposed technique consists of two steps: first, the Eikonal equations describing the geometric-optic ray paths are solved using an efficient Runge-Kutta routine to construct a set of possible photon migration paths through the inhomogeneous tissue sample. Thereafter, the transient photon transport equation, written in terms of the arc length, is solved along each ray path to construct the optical intensity variation as time progresses. The main advantage of the current algorithm is that it enables to discretise the pulse propagation space adaptively by taking optical depth into account. Therefore, computational efficiency can be increased without compromising the accuracy of the algorithm. Computational efficiency becomes a bottle-neck when large, realistic simulations of optical pulse propagation in tissue are required. Therefore, current proposed method is very much suited for extensive computations work in time-resolved photon-diffusion tomography work.

## REFERENCES

[1] S. B. Colak, M. B. van der Mark, G. W. 'T Hooft, J. H. Hoogenraad, E. S. van der Linden, and F. A. Kuijpers, "Clinical optical tomography and NIR spectroscopy for breast cancer detection," *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 5, no. 4, pp. 1143–1158, 1999.

[2] B. D. Cameron, H. W. Gorde, B. Satheesan, and G. L. Cote, "The use of polarized laser light through the eye for noninvasive glucose monitoring," *Diabetes Technology & Therapeutics*, vol. 1, no. 2, pp. 135–143, 1999.

[3] F. Liu, K. M. Yoo, and R. R. Alfano, "Ultrafast laser-pulse transmission and imaging through biological tissues," *Applied Optics*, vol. 32, no. 4, pp. 554–558, 1993.

[4] G. Yao and L. V. Wang, "Theoretical and experimental studies of ultrasound-modulated optical tomography in biological tissue," *Applied Optics*, vol. 39, no. 4, pp. 659–664, 2000.

[5] A. D. Klose, U. Netz, J. Beuthan, and A. H. Hielscher, "Optical tomography using the time-dependent equation of radiative transfer—part 1: forward model," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 72, no. 5, pp. 691–713, 2002.

[6] M. Premaratne, E. Premaratne, and A. J. Lowery, "The photon transport equation for turbid biological media with spatially varying isotropic refractive index," *Optics Express*, vol. 13, no. 2, pp. 389–399, 2005.

[7] K. Stamnes, S. Tsay, W. Wiscombe, and K. Jayaweera, "Numerically stable algorithm for discrete-ordinate-method for radiative transfer in multiple scattering and emitting layered media," *Applied Optics*, vol. 27, no. 12, pp. 2502–2509, 1988.

[8] A. D. Kim and M. Moscoso, "Chebyshev spectral methods for radiative transfer," *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 2074–2094, 2002.

[9] J. V. P. de Oliveira, A. V. Cardona, and M. T. M. B. de Vilhena, "Solution of the one-dimensional time-dependent discrete ordinates problem in a slab by the spectral and LTS$_N$ methods," *Annals of Nuclear Energy*, vol. 29, no. 1, pp. 13–20, 2002.

[10] M. Sakami, K. Mitra, and P.-F. Hsu, "Analysis of light pulse transport through two-dimensional scattering and absorbing media," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 73, no. 2–5, pp. 169–179, 2002.

[11] T. Khan and H. Jiang, "A new diffusion approximation to the radiative transfer equation for scattering media with spatially varying refractive indices," *Journal of Optics A: Pure and Applied Optics*, vol. 5, no. 2, pp. 137–141, 2003.

[12] H. Dehghani, B. Brooksby, K. Vishwanath, B. W. Pogue, and K. D. Paulsen, "The effects of internal refractive index variation in near-infrared optical tomography: a finite element modelling approach," *Physics in Medicine and Biology*, vol. 48, no. 16, pp. 2713–2727, 2003.

[13] G. Bal, "Radiative transfer equations with varying refractive index: a mathematical perspective," *Journal of the Optical Society of America A*, vol. 23, no. 7, pp. 1639–1644, 2006.

[14] M. L. Shendeleva and J. A. Molly, "Scaling property of the diffusion equation for light in a turbid medium with varying refractive index," *Journal of the Optical Society of America A*, vol. 24, no. 9, pp. 2902–2910, 2007.

[15] L. Martí-López, J. Bouza-Domínguez, R. A. Martínez-Celorio, and J. C. Hebden, "An investigation of the ability of modified radiative transfer equations to accommodate laws of geometrical optics," *Optics Communications*, vol. 266, no. 1, pp. 44–49, 2006.

[16] A. Sharma, D. V. Kumar, and A. K. Ghatak, "Tracing rays through graded-index media: a new method," *Applied Optics*, vol. 21, no. 6, pp. 984–987, 1982.

[17] B. Richerzhagen, "Finite element ray tracing: a new method for ray tracing in gradient-index media," *Applied Optics*, vol. 35, no. 31, pp. 6186–6189, 1996.

[18] H. A. Ferwerda, "Radiative transfer equation for scattering media with a spatially varying refractive index," *Journal of Optics A: Pure and Applied Optics*, vol. 1, no. 3, pp. L1–L2, 1999.

[19] C. C. Handapangoda, M. Premaratne, L. Yeo, and J. Friend, "Laguerre Runge-Kutta-Fehlberg method for simulating laser pulse propagation in biological tissue," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 14, no. 1, 2008.

[20] M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, Cambridge, UK, 7th edition, 1999.

[21] M. Premaratne, "Numerical simulation of nonuniformly time-sampled pulse propagation in nonlinear fiber," *Journal of Lightwave Technology*, vol. 23, no. 8, pp. 2434–2442, 2005.

[22] A. D. Greenwood and J.-M. Jin, "A field picture of wave propagation in inhomogeneous dielectric lenses," *IEEE Antennas and Propagation Magazine*, vol. 41, no. 5, pp. 9–18, 1999.

[23] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Integration of functions," in *Numerical Recipes in C++*, pp. 152–157, Cambridge University Press, Cambridge, UK, 2003.

[24] M. Abramomitz and I. A. Stegun, "Orthogonal polynomials," in *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, pp. 773–784, Dover, New York, NY, USA, 1965.

[25] S. C. Chapra and R. P. Canale, "Runge-Kutta methods," in *Numerical Methods for Engineers*, pp. 719–720, McGraw-Hill, New York, NY, USA, 4th edition, 2002.

[26] M. P. Mengüç and R. Viskanta, "Radiative transfer in three-dimensional rectangular enclosures containing inhomogeneous, anisotropically scattering media," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 33, no. 6, pp. 533–549, 1985.

[27] G. E. Thomas and K. Stamnes, *Radiative Transfer in the Atmosphere and Ocean*, Cambridge University Press, Cambridge, UK, 1999.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 369651, 11 pages
doi:10.1155/2008/369651

*Research Article*

# Haptic Stylus and Empirical Studies on Braille, Button, and Texture Display

**Ki-Uk Kyung, Jun-Young Lee, and Junseok Park**

*POST-PC Research Group, Electronics and Telecommunications Research Institute, Yuseong-Gu, Daejeon 305-700, South Korea*

Correspondence should be addressed to Ki-Uk Kyung, kyungku@etri.re.kr

Received 2 September 2007; Revised 16 December 2007; Accepted 24 December 2007

Recommended by Daniel Howard

This paper presents a haptic stylus interface with a built-in compact tactile display module and an impact module as well as empirical studies on Braille, button, and texture display. We describe preliminary evaluations verifying the tactile display's performance indicating that it can satisfactorily represent Braille numbers for both the normal and the blind. In order to prove haptic feedback capability of the stylus, an experiment providing impact feedback mimicking the click of a button has been conducted. Since the developed device is small enough to be attached to a force feedback device, its applicability to combined force and tactile feedback display in a pen-held haptic device is also investigated. The handle of pen-held haptic interface was replaced by the pen-like interface to add tactile feedback capability to the device. Since the system provides combination of force, tactile and impact feedback, three haptic representation methods for texture display have been compared on surface with 3 texture groups which differ in direction, groove width, and shape. In addition, we evaluate its capacity to support touch screen operations by providing tactile sensations when a user rubs against an image displayed on a monitor.

## 1. INTRODUCTION

Researchers have proposed a diverse range of haptic interfaces for more realistic communication methods with computers. Force feedback devices, which have attracted the most attention with their capacity to physically push and pull a user's body, have been applied to game interfaces, medical simulators, training simulators, and interactive design software, among other domains [1]. However, compared to force feedback interfaces, tactile displays have not been deeply studied. It is clear that haptic applications for mobile devices, such as PDAs, mobile computers, and mobile phones, will have to rely on tactile devices. Such a handheld haptic system will only be achieved through the development of a fast, strong, small, silent, safe tactile display module, with low heat dissipation and power consumption. Furthermore, stimulation methods reflecting human tactile perception characteristics should be suggested together with a device.

A number of researchers have proposed tactile display systems. In order to provide tactile sensation to the skin, work has looked at mechanical, electrical, and thermal stimulation. Most mechanical methods involve an array of pins driven by linear actuation mechanisms such as solenoids, piezoelectric actuators, or pneumatic actuators. An example is the "Texture Explorer," developed by Ikei and Shiratori [2]. This $2\times5$ flat pin array is composed of piezoelectric actuators and operates at a fixed frequency (~250 Hz) with maximum amplitude of 22 $\mu$m. Summers and Chanter [3] developed a broadband tactile array using piezoelectric bimorphs and reported empirical results for stimulation frequencies of 40 Hz and 320 Hz with the maximum displacement of 50 $\mu$m. Since the aforementioned tactile displays may not result in sufficiently deep skin indentation, Kyung et al. [4] developed a $5 \times 6$ pin-array tactile display which has a small size, long travel, and high bandwidth. However, this system requires a high input voltage and a high power controller. As an alternative to providing normal indentation, Hayward and Cruz-Hernandez [5] and Luk et al. [6] have focused on the tactile sensation of lateral skin stretch and designed a tactile display device which operates by displaying distributed lateral skin stretch at frequencies of up to several kilohertz. However, it is arguable that the device remains too large (and high voltage) to be realistically integrated into a mobile device. Furthermore, despite work investigating user performance on cues delivered by lateral skin stretch, it remains unclear whether this method is capable of displaying the full range of stimuli achievable by presenting an array of normal forces.

Konyo et al. [7] used an electroactive polymer as an actuator for mechanical stimulation. Poletto and Van Doren developed a high-voltage electrocutaneous stimulator with small electrodes [8]. Kajimoto et al. [9] developed a nerve axon model based on the properties of human skin and proposed an electrocutaneous display using anodic and cathodic current stimulation. Unfortunately, these tactile display devices sometimes involve user discomfort and even pain.

We can imagine a haptic device providing both force and tactile feedback simultaneously. Since Kontarinis and Howe applied vibration feedback to a teleoperation in 1995 [10], some research works have had interests in combination of force and tactile feedback. Akamatsu and MacKenzie [11] suggested a computer mouse with tactile- and force feedback-increased usability. However, the work dealt with haptic effects rather than precisely controlled force and tactile stimuli. In 2004, Kammermeier et al. combined a tactile actuator array providing spatially distributed tactile shape display on a single fingertip with a single-fingered kinesthetic display and verified its usability [12]. However, the size of the tactile display was not small enough to practically use the suggested mechanism. As more practical design, Okamura et al. designed a 2D tactile slip display and installed it into the handle of a force feedback device [13]. Recently, in order to provide texture sensation with precisely controlled force feedback, a mouse fixed on 2DOF mechanism was suggested [14]. A small pin-array tactile display was embedded into a mouse body and it realized texture display with force feedback. More recently, Allerkamp et al. developed a compact pin-array and they tried to realize the combination of force feedback and tactile display based on the display and vibrations [15]. However, in previous works, the tactile display itself is quite small but its power controller is too big to be used practically. Our work in this paper deals with this issue as one of applications of our system.

In the area of human tactile perception, Johansson and Vallbo [16] and Johnson and Phillips [17] have studied human mechanoreceptors and their function in connection with tactile perception and the anatomical structure of glabrous skin such as the palm or finger pad. Verrillo et al. have suggested a four-channel model of vibrotaction which shows the variation of the displacement (indentation depth) threshold to frequency [18]. Also, studies have measured the sensation magnitude of thresholds as a function of frequency of vibration [18, 19]. The previous physiological research shows that humans have four types of mechanoreceptors for tactile sense and that each type responds in a specific band of frequency. Therefore, frequency characteristics should be given careful consideration in the design of a tactile display device and stimulation method.

In this paper, we propose a compact tactile display module which can be embedded into small devices and a pen-type haptic interface providing impact and distributed pressure. In Section 2, the design parameters and structure of the proposed tactile display module are described in detail. In Section 3, the implementation of a pen-like haptic interface including the tactile display module and impact generator is presented. In Section 4, we evaluate performance of this system, which we term the "Ubi-Pen II." In Section 5, performance of a force and tactile feedback interface adopting the suggested pen-like interface is described. Finally, in Section 6, we discuss possible applications of the proposed system including image display on a touch screen.

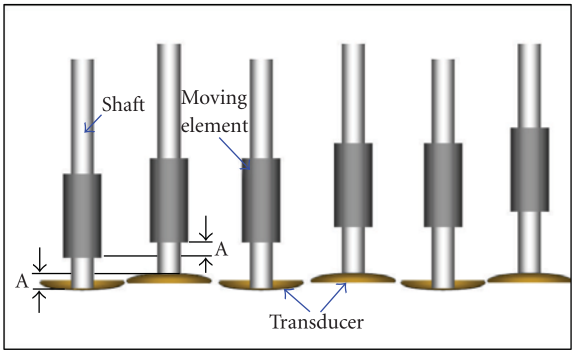


Figure 1: Operation principle of an actuator.

## 2. COMPACT TACTILE DISPLAY MODULE

### 2.1. *Design of a tactile display module*

In order to make a tactile display module, actuator selection is the first and dominant step. The actuator should be small, light, safe, silent, fast, powerful consume modest amounts of power and emits little heat. Recently, we developed a small tactile display using a small ultrasonic linear motor [20]. We here briefly describe its operation principle and mechanism.

The basic structure and driving principle of the actuator are described in Figure 1. The actuator is composed of a transducer, a shaft, and a moving element. The transducer is composed of two piezoelectric ceramic disks and elastic material membranes. The convex motion of the membranes causes lift in the shaft of the motor. The fast restoring concave motion overcomes the static frictional force between the moving element and the shaft, and it makes the moving element maintain its position. The displacement "A" of one cycle is submicrometer scale, and the rapid vibration of the membrane at a frequency of 45 kHz (ultrasonic range) causes rapid movement of the moving element. The diameter of the transducer is 4 mm and its thickness is 0.5 mm. The thrusting force of the actuator is greater than 0.2 N and the maximum speed of the moving element is around 30 mm/sec. In order to minimize the size of the tactile display module, the actuators were arranged as shown in Figure 2. Essentially, this figure shows the arrangement of two variations on the actuators—each with different shaft lengths. This design minimizes the gap between actuators. Another feature is that the elements previously described as "moving" are now stationary and fixed together, causing the shafts to become the elements which move when the actuators are turned on. This minimizes the size of the contact point with a user's skin (to the 1 mm diameter of the shaft), while maintaining the mechanical simplicity of the system.

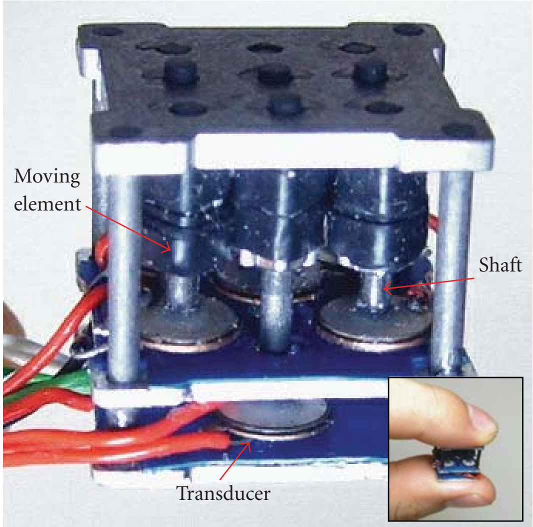


FIGURE 2: The implemented tactile display module.

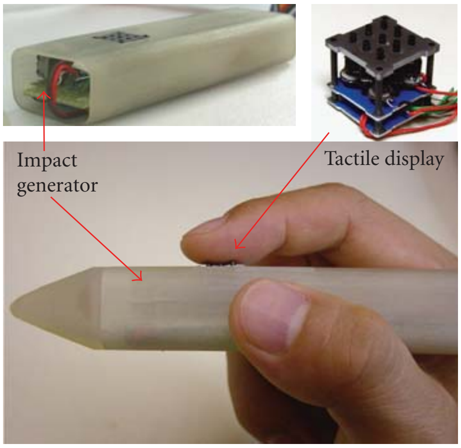


FIGURE 3: The prototype of the Ubi-Pen II.

### 2.2. Implementation

From the design specification described in Section 2.1, the prototype of the tactile display module has been implemented as shown in Figure 2. In order to embed the module in a pen, we constructed only a 3 × 3 pin array. However, it should be noted that the basic design concept is fully extensible; additional columns and rows can be added without electrical interference or changes in pin density. The shaft itself plays the role of tactor and has a travel of 1 mm. The distance between two tactors is 3.0 mm. Since the actuators operate in the ultrasonic range, they produce little audible noise. The average thrusting force of each actuator exceeds 0.2 N, sufficient to deform the skin with an indentation of 1 mm [21]. The total size of the module is 12 × 12 × 12 mm and its weight is 2.5 g. Since the maximum speed of a pin is around 30 mm/sec, the bandwidth of the tactile display is approximately 20 Hz when used with a maximum normal displacement of 1 mm. If the normal displacement is lower than 1 mm, the bandwidth could be increased.

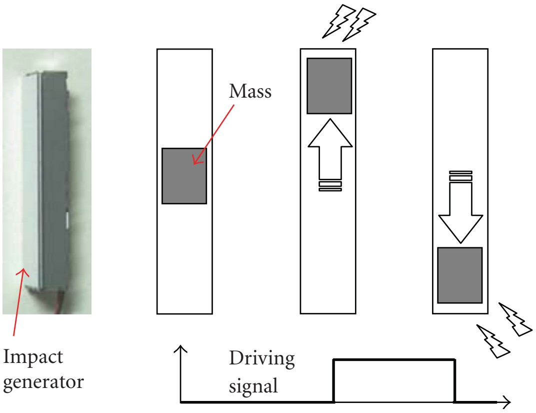


FIGURE 4: Operation principle of an impact generator.

## 3. IMPLEMENTATION OF HAPTIC STYLUS

The styli have become common tools for interacting with mobile communication devices. In the area of haptics, Lee et al. [22] suggested a haptic pen which could provide a sense of contact based around a touch sensor and a solenoid. It could generate a feeling corresponding to clicking a button.

In order to support richer stylus-based tactile cues, we embedded our tactile display module into a pen-like prototype. We termed these kinds of devices the Ubi-Pen and intend it for use as an interface to VR, for the blind, to represent textures, and as a symbolic secure communication device [20]. In our previous version, a small vibrator was installed at the tip of the pen. However, since the vibrator's temporal response is slow, it causes time delay between signal and activation. Although it was effective, it was not realistic.

In this research, instead of a typical vibrator, we installed an impact generator in the head of the pen to provide a sense of contact (see Figure 3). We named this version the Ubi-Pen II. We suggest that it could be used generally as the stylus of a mobile communication device, which provides realistic and interactive haptic cues such as buttons during operation of OS.

Figure 4 shows an operation principle of the impact generator. There is a mass inside the generator and electromagnetic force induced by electric signal that makes the mass move along a longitudinal axis of the case. This generator is generally used as a kind of linear vibrator and we otherwise use it as an impact generator. The generator is arranged along a longitudinal axis of the stylus housing. When a rising signal is applied to the generator, the mass moves up fast and it collides with the upper side. When a falling signal is applied to the generator, the mass moves down fast and it collides with the bottom side. The response time of the mass movement is within milliseconds scale.

## 4. EVALUATION OF PERFORMANCE

### 4.1. Braille display of the tactile display module

A common method to evaluate the performance of tactile displays is to test user's performance at recognizing specific

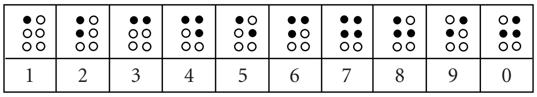


Figure 5: Braille patterns for the experiment.

Table 1: Experimental results.

| | Normal subjects | Blind subjects |
|---|---|---|
| Average percentage of correct answers | 80.83 | 100 |
| Average duration of each trial (sec) | 5.24 | 1 ∼ 2 |

patterns [2, 4]. We use Braille as a stimulus set to conduct such a test. Specifically, we conducted a study involving the presentation of the Braille numbers 0 ∼ 9 on the Ubi-Pen.

Figure 5 shows the experimental Braille patterns. Subjects were required to hold the pen such that the tip of their index finger rested over the pin-array part of tactile display module. In our previous work, the test was conducted for the normal people and there was small observations for the blind [20]. In this paper, the Braille display test bas been conducted for the normal and the blind.

After setup stage, we conducted a study on recognition rate of the 10 numeric digits in the Braille character set. As these can be displayed on only four pins, we mapped them to the corner pins on our tactile display module. We chose to do this as our user-base was composed of sighted Braille novices. We used three different stimulation frequencies: 0, 2, and 5 Hz. (Pins move up and maintain static position at the 0 Hz.) Pins movement was synchronized. We presented 60 trials in total, each number at each frequency, twice. All presentations were in a random order, and subjects were not advised about the correctness of their responses. 10 subjects participated in the experiment. The Braille stimuli were generated continuously and changed as soon as the subject respond using the graphic user interface. There were 2-minute breaks after every 20 trials.

Two blind people have participated in the same experiment and the visual guidance in the experiment has been replaced by the speech guidance of experimenter. For all stimuli, they responded exactly and quickly. The Braille expert usually read more than 100 characters [23], and the blind subjects responded that they do not feel any difficulties to read the Braille numbers. Since the duration of each trial was shorter than 1 ∼ 2 seconds and they answer in the form of speech, we could not measure the duration exactly.

Moreover, 4 neighborhood pins have been presented again with identical procedure for the blind people; and they responded more quickly since the gap of pins was more familiar with them. Duration of each trial was always shorter than 1 second.

Table 1 shows the summary of experimental results. Although normal subjects were novice in using the tactile display, the average percentage of correct answers exceeded 80 percent. The confusions come from the relatively low tactile sensitivity of the novices compared with the sensitivity of the blind. Since the various analysis of the tactile display for the blind is another interesting topic, this will be investigated in our future work.

Craig's research shows the blind people have extraordinary capability to recognize the vibrotactile patterns at very high frequencies [23]. It might be true that specialized people recognize vibrotactile patterns without respect to frequencies. However, spatial acuity of human tactile perception is a function of the vibration frequency; and we need to determine the best frequency for the tactile pattern display using the developed device. Our previous work shows spatial acuities are better at the range of the Merkel's disk and Meisner's corpuscle [4]. From the comparisons at the frequency range of 0 ∼ 560 Hz, the sensitive range of the Merkel's disk, 1 ∼ 3 Hz, was the best frequency for the pattern perception since the mechanoreceptor is mainly related to the sense of surface pattern and distributed pressure [18]. Before conducting the experiment, we needed to look at the frequency bands of peripheral tactile neural responses. There are four mechanoreceptors in the glabrous skin of the palm and fingertip regions. Meissner's corpuscles and Merkel's discs are located in the upper layers, and Ruffini endings and Pacinian corpuscles are located more deeply. These receptors are divided into the following two classes according to their rate of adaptation: the slowly adapting afferent receptors and the rapidly adapting afferent receptors. The slowly adapting afferent receptors comprise Merkel's discs (SA I) and the Ruffini endings (SA II), while the rapidly adapting afferent receptors comprise Meissner's corpuscles (RA I) and the Pacinian corpuscles (RA II). The four mechanoreceptors each have different functions [16, 18]. The SA I afferents respond to quasistatic deformations of the skin, such as force or displacement in the frequency range of 0.4–3 Hz. These receptors play an important role in detecting spatial structures in static contact, such as an edge or a bar. The size of Merkel's receptor is small and shows very high innervation density at the tip of index finger. The SA II afferent receptors provide a neural image related to the direction of the skin being stretched. SA Type II fibers produce a buzz-like sensation in the frequency range of 100–500 Hz. The RA I afferent receptors, which have a frequency range of 2–40 Hz, detect dynamic deformations of the skin such as the sensation of flutter. The RA I afferent receptors are about four times more sensitive than the SA I afferent receptors; in addition, RA I shows best sensitivity in the frequency range of 25–40 Hz. The RA II afferent receptors, which have a frequency response in the range of 40–500 Hz, are the most sensitive to vibration amplitude and are particularly known to serve as detectors of acceleration or vibration. Previous anatomic study shows the size of Pacinian corpuscles to be bigger than the other mechanoreceptors located deeper within the skins, and their innervation density is low [24]. Therefore, it is to be expected that their spatial acuity would be poor. (However, in some cases [23], good spatial resolution may be observed at frequencies expected to activate Pacinian corpuscles.) Based on these findings, we found that humans were more sensitive at a frequency band of 1 ∼ 3 Hz in tactile pattern discrimination that they are at surrounding frequencies [4]. This is due to the structure of our neural mechanism for sensing tactile

Table 2: Percentage of correct answers according to frequencies.

| | 0 Hz | 2 Hz | 5 Hz |
|---|---|---|---|
| Average percentage of correct answers | 79.9 | 81.9 | 82.7 |
| Standard deviation | 18.6 | 12.3 | 9.2 |

pattern. One part is easily activated by this frequency band. Therefore, we hypothesized that stimuli delivered in that frequency range would outperform those outside it. This was brought out by asking subjects about their impressions of the cues, and 8 of the 10 subjects suggested that some frequencies were easier to detect than others.

However, as shown in Table 2, there is no difference among the percentage of correct answers according to frequencies. Investigating in more detail, we turned to task completion time. Average duration of a trial was 5.98 seconds at the 0 Hz, 4.42 seconds at the 2 Hz, and 5.24 seconds at 5 Hz. Thus, the average duration of a trial is decreased at the 2-Hz frequency. Although, inconclusive, we suggest this indicates that subjects found the sensations delivered at this frequency to be easier to detect. In this section, the performance of the tactile display module has been verified. Especially, its capability of displaying Braille for the blind was proved. In addition, an appropriate stimulating frequency has been investigated.

Here, we have some issues to be discussed. As mentioned previously, since the blind people are familiar with rubbing surface to read the Braille, we are not sure that stimulation of 2 Hz is effective for the blind. In fact, after they participated in the experiments, they commented that static display was easier to discriminate than vibrational stimuli. We have to consider user's familiarity when we design tactile stimuli.

### 4.2. *Simulation of button pressing sense*

One of the most frequent complaints when using a touch screen is ambiguity about whether a screen tap has resulted in a successful button press. Researchers have proposed that there is a touch screen providing active touch feedback to address this issue [25]. In a previous version of the Ubi-Pen, there is a short-term vibration feedback for notifying button clicking [20]. In a different manner, the Ubi-Pen II also possesses the ability to produce a click-like sensation with an impact generator.

As shown in Figure 6, button pressing is composed of 3 steps. The first step is increasing pressing force. The second step is button pressed state after sudden falling down when the pressing force is greater than a threshold. The third step is releasing the button with an abrupt rising up. We do not have to consider the first step since it naturally occurs on a touch screen. The touch screen itself provides a function of button pressing with a threshold pressure; and the keys of the second and the third steps are sudden change of movement. Because the sudden change is a kind of impact, we can simulate the second and the third steps with our haptic stylus including an impact generator. As shown in Figure 6, the falling down collision of the mass inside the generator gives effect of the button pressing. The rising up collision of the mass provides sense of the button releasing to users.

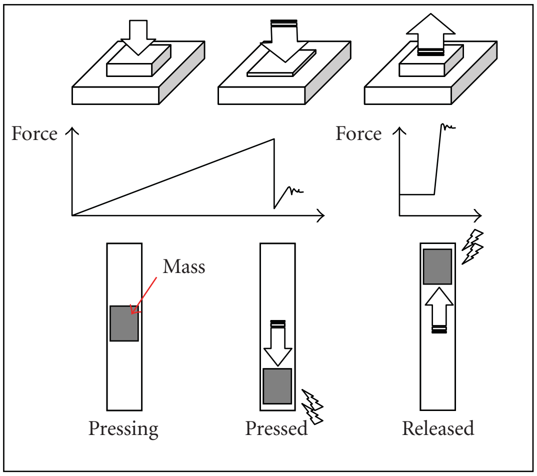
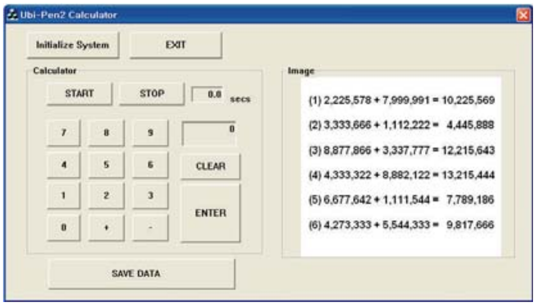


Figure 6: Procedure of button pressing sense.



(1) 2,225,578 + 7,999,991 = 10,225,569
(2) 3,333,666 + 1,112,222 = 4,445,888
(3) 8,877,866 + 3,337,777 = 12,215,643
(4) 4,333,322 + 8,882,122 = 13,215,444
(5) 6,677,642 + 1,111,544 = 7,789,186
(6) 4,273,333 + 5,544,333 = 9,817,666

Figure 7: Calculator and presented equations.

Table 3: Effectiveness button pressing sense feedback.

| | Average duration of calculation | Standard deviation |
|---|---|---|
| Without haptic feedback | 14.04 (sec) | 2.62 |
| With haptic feedback | 10.66 (sec) | 2.15 |

Here we test the effectiveness of this feature. We presented subjects with a simple calculator interface, shown in Figure 7. They had to enter each of the 6 equations shown on the right of the screen. Each equation was randomly presented and haptic feedback was also randomly provided in half the trials. Subjects had to calculate every equation twice until they obtained the correct answer to each. This calculator displayed only the results of calculations, not the figures entered. In this study, we measured task completion time

The experimental results in Table 3 show that the clicking sense feedback of the Ubi-Pen II decreased the length of time to enter the calculations. The major influence of the click sensation was to add self confidence to users, and this contributed to the production of fewer errors and the reduced duration of the calculations. We asked each participant about the effectiveness of clicking sense feedback and they all agreed that clicking sense feedback gives self confidence and reality.

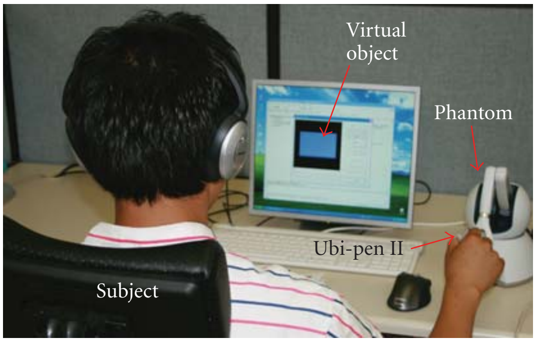


Figure 8: Force and tactile feedback interface.

Additionally, we had a chance demonstrating the Ubi-Pen II at an IT exhibition show and 145 of 160 visitors agreed that proposed scheme provide users with reality of a button. From this test, the effectiveness of the Ubi-Pen's button pressing feedback has been verified.

## 5. COMBINATION OF FORCE FEEDBACK AND TEXTURE FEEDBACK

### 5.1. *System and experimental design*

Currently, the PHANToM is the most widely used haptic interface. It has force feedback capabilities and it provides a stylus-like handle interface [26]. Here we replace its handle with the Ubi-Pen II to add tactile feedback capability to the device. Since the Ubi-Pen provides both impact and texture stimuli, this allows us to compare the effectiveness of various haptic stimulation methods.

In our previous experiment, the previous version of the Ubi-Pen provided texture feedback and vibration feedback [20]. However, we reported that vibration potentially had problems in aspect of control. The stylus is replaced by the Ubi-Pen II in this experiment. We conduct similar experiment here, but we observe the effectiveness of impact feedback on texture display. As shown in Figure 8, the proposed pen-like interface was attached to the handle of a force feedback device (model: PHANToM Omni). In order to test performance of the system, we designed a virtual tangible object. The virtual object is a box and its stiffness is 2 kN/m. (The task in this experiment does not require high interaction force.) The widths are 75 mm (300 pixels) and 67.5 mm (270 pixels). The upper surface of the box has a texture derived from texture mapping an image and a user explores only upper surface. In order to use the image as a texture, this test provides a symbolic pointer in the shape of a square, with a size of 15 × 15 pixels. A user can load any gray-scale image. As shown in Figure 9, when the user touches an image on the box with the integrated interface, the area of the cursor is divided into 9(= 3 × 3) subcells and the average gray value of each cell is calculated. Then, this averaged gray value is converted to the intensity of the stimuli displayed on each pin of the tactile display.

In this interaction, the stiffness of the box is represented by the PHANToM force feedback device. However, the texture on the surface can be represented in 3 ways. The first is through force feedback presented by the PHANToM since we can feel texture by probe scanning. The second is texture feedback by the Ubi-Pen since the pin's movement can display surface roughness. The third is the Ubi-Pen's impact feedback since such stimuli could facilitate the recognition of obstacles when rubbing a surface. We compared all the 3 possible stimulation methods in this experiment as shown in Figure 9. As mentioned above, the area of virtual cursor is divided into 9 cells each with an individual gray value. However, while the tactile display inside the pen interface has 9 spatially distributed stimulators, the impact and force feedback interface both have only one interaction point. Therefore, force feedback and impact feedback use only the center value.

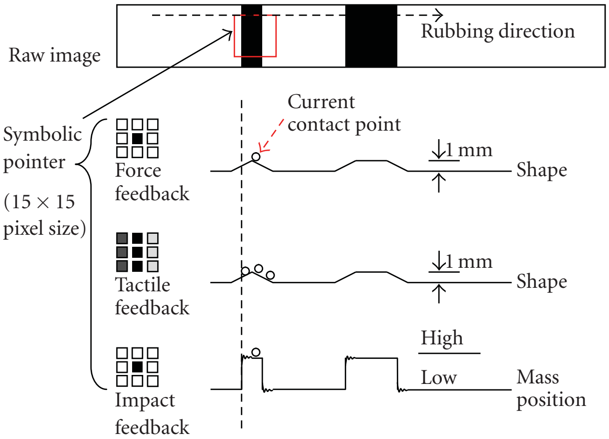


Figure 9: Methodology of texture display according to the stimulation method.

In case of force feedback, the gray value is converted into the height of pattern and its highest value is 1 mm. In case of tactile feedback, the gray value is converted into the normal displacement of each pin and the maximum displacement is 1 mm. When we use a pin-array-based tactile display, representing resolution of the tactile display is determined by the resolution of the pin-array. Thus, only tactile display with high density pin-array is the solution of the high-resolution display. In order to make up this limitation, we derived an idea that the tactile display plays a role of a texture magnifier. As shown in Figure 10, size of the tactile display is 2.4 times bigger than the symbolic pointer. This kind of skill may decrease reality in aspect of size, but it is a useful tip to convey texture information to a user precisely when we use a low-density pin-array.

In case of impact feedback, haptic cues indicate change of region while the pointer across over the texture pattern. When the pointer moves inside texture area, the mass rises up and a user recognizes a ridge of the pattern. When the pointer escapes texture area and the gray value decreases under a threshold value, the mass falls down and the user experiences sudden drop-like feeling. This kind of stimulation may not precisely represent projected shapes of textures that could be effective to display surface patterns.

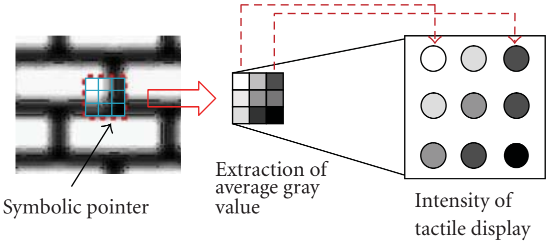
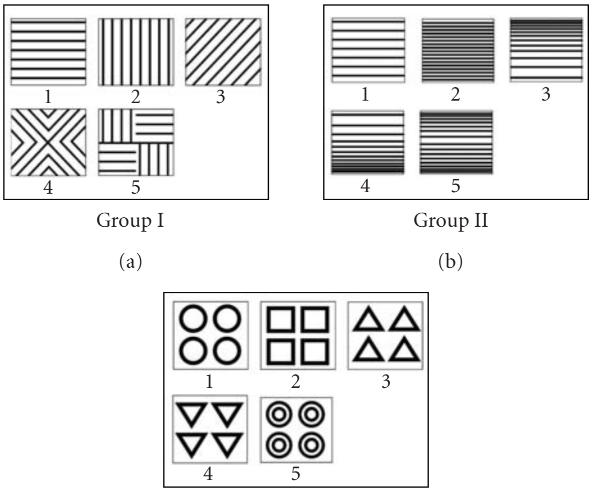

Symbolic pointer
Extraction of average gray value
Intensity of tactile display

Figure 10: Methodology of pattern display.



Group I

(a)

Group II

(b)

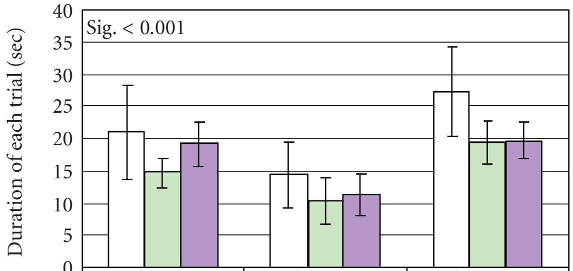
Group III

(c)

Figure 11: Texture samples.

In order to compare the performance of all stimulation methods, we prepared 3 groups of tactile patterns. Figure 11(a) shows 5 image samples from group I which differ in the direction of the gratings they feature. The size of each image was 300 × 270 pixels. Figure 11(b) shows image samples from group II which contains grooves of varying widths. A user feels horizontal gratings while rubbing the surfaces. In order to discriminate these patterns, the tactile stimuli must be integrated with movements on the plane. Figure 11(c) shows 5 image samples from group III, each of which shows different shapes. Discriminating among these patterns will require precise and accurate integration of the tactile cues with the movements on the surface. Feeling distributed pressure (as with the pin array display) may help users to discern the surfaces.

Ten subjects participated in the experiment. In each trial, one of the five images from one of the groups was texture mapped on the upper surface of a virtual box. However, the graphical representation was hidden, and only a blank surface displayed. When the user touched and rubbed the surface of the object, the gray values of the image were conveyed to the haptic interface. They were then required to state which texture was present. The subjects have shown all images patterns through another screen in order to make their choice. All texture images in a group were presented 4 times



| | Group I | Group II | Group III |
|---|---|---|---|
| □FF | 21 | 14.5 | 27.3 |
| 🟩TF | 14.7 | 10.4 | 19.3 |
| 🟪IF | 19.1 | 11.1 | 19.6 |

Figure 12: Duration of each trial.

at random and the order of test group was also randomly selected. The user felt the stiffness of the box by force feedback, but there were three conditions for representing texture: force feedback, tactile feedback, and impact feedback. In order to prevent practice effects, the order of the stimulation method was also randomized. Finally, sounds produced during the interaction may affect recognition performance, so participants were required to wear noise cancelling headphones (Bose, QuietComfort2).

### 5.2. *Performance and discussion*

Table 4 shows experimental results for the force feedback case in the form of a confusion matrix. Likewise, Tables 5 and 6, respectively, show the experimental results for tactile and impact feedback. In case of force feedback, average percentages of correct answers are 86.5% for group I, 73.5% for group II, and 60.5% for group III. In case of tactile feedback, average percentages of correct answers are 97.5% for group I, 91.5% for group II, and 80.5% for group III. In case of impact feedback, average percentages of correct answers are 83.5% for group I, 81.5% for group II, and 61.0% for group III. Figure 12 shows the mean durations of trials in each condition. The experimental results for force feedback and tactile feedback are similar to the previous paper's results [20]. This confirms that both previous and new experimental results are reliable. In case of impact feedback, since impact plays a role of cue to notifying change of texture, experimental results are a bit similar to the case of vibration feedback previously observed.

The texture samples assigned in group I can be discriminated by detecting the direction of the gratings. Users can recognize the direction from the position of the interaction point and the direction in which they rub. In this case, there is no substantial difference between force feedback and impact feedback. However, tactile display provides line load to the finger along the gratings. As shown in Tables 4, 5, and 6 as well as Figure 12, this makes human recognize direction of the gratings more correctly and quickly.

For group II, the images can be discriminated by the variations in the spacing between the ridges. However, the spatial resolution of the human arm is not sufficient to reliably detect variations on the scale of millimeters whereas the skin

Table 4: Experimental results for force feedback (%).

| Force feedback | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Group I | 1 | **95.0** | 2.5 | — | 2.5 | — |
| | 2 | — | **75.0** | 5.0 | 12.5 | 7.5 |
| | 3 | 7.5 | 5.0 | **85.0** | 2.5 | — |
| | 4 | — | — | 5.0 | **95.0** | — |
| | 5 | 15.0 | 2.5 | — | — | **82.5** |
| Group II | 1 | **82.5** | 2.5 | 7.5 | 7.5 | — |
| | 2 | 2.5 | **67.5** | — | 12.5 | 17.5 |
| | 3 | 12.5 | 10.0 | **75.0** | — | 2.5 |
| | 4 | — | 12.5 | — | **82.5** | 5.0 |
| | 5 | 2.5 | 20.0 | 5.0 | 12.5 | **60.0** |
| Group III | 1 | **55.0** | 15.0 | 12.5 | 17.5 | — |
| | 2 | 22.5 | **60.0** | 15.0 | — | 2.5 |
| | 3 | 25.0 | 7.5 | **55.0** | 12.5 | — |
| | 4 | 7.5 | 5.0 | 10.0 | **67.5** | 10.0 |
| | 5 | 7.5 | 15.0 | — | 12.5 | **65.0** |

Table 5: Experimental results for tactile feedback (%).

| Tactile feedback | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Group I | 1 | **100.0** | — | — | — | — |
| | 2 | — | **100.0** | — | — | — |
| | 3 | — | — | **97.5** | 2.5 | — |
| | 4 | — | — | 7.5 | **92.5** | — |
| | 5 | — | — | — | 2.5 | **97.5** |
| Group II | 1 | **95.0** | — | 2.5 | 2.5 | — |
| | 2 | — | **100.0** | — | — | — |
| | 3 | — | 7.5 | **92.5** | — | — |
| | 4 | — | — | — | **97.5** | 2.5 |
| | 5 | — | 22.5 | — | 5.0 | **72.5** |
| Group III | 1 | **60.0** | 17.5 | 12.5 | 10.0 | — |
| | 2 | 10.0 | **90.0** | — | — | — |
| | 3 | 5.0 | — | **95.0** | — | — |
| | 4 | — | — | 7.5 | **82.5** | 10.0 |
| | 5 | 10.0 | — | — | 15.0 | **75.0** |

sense allows discrimination of submillimeter gaps [17]. In addition, pattern display by force feedback inherently results in movement of the arm and even stick slip vibration, factors which may disturb discrimination of gap variation. Therefore, as shown in Table 4, the percentage of correct answers for force feedback is lower than in the other conditions. A good example is that users experienced difficulty discriminating between sample 2 and sample 5. In the case of the tactile feedback, the narrow gaps are discriminated though the skin. This shows the best performance. In the case of the impact feedback, the participants typically rubbed the surface at a constant speed and felt the frequency of the stimulation. This technique was also effective.

As mentioned in Section 5.1, in order to recognize shape of a pattern, the tactile stimuli must be accurately integrated with movements on the plane. However, arm movements do not guarantee the high spatial resolution required for this. For example, when sample 3 of group III was presented, users found it hard to discern it from the other samples; but, in case of the tactile feedback, the distributed pressure cues enabled them to make more accurate choices.

If the tactile display had more pins, it might show better performance. However, over all the tests, the haptic device combined with the built-in compact tactile display showed satisfactory results. Impact feedback was also reasonably effective in texture display with force feedback.

## 6. APPLICATION OF THE Ubi-Pen II

### *6.1. Image display on touch screen*

As shown in Figure 13, the Ubi-pen mouse enables tactile pattern display when the scheme described in Section 5.1 is applied to the image on a touch screen. In order to verify

Table 6: Experimental results for impact feedback (%).

| Impact feedback | | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 |
|---|---|---|---|---|---|---|
| Group I | 1 | **85.0** | — | 5.0 | — | 10.0 |
| | 2 | 5.0 | **90.0** | 5.0 | — | — |
| | 3 | — | — | **85.0** | 10.0 | 5.0 |
| | 4 | — | 10.0 | 15.0 | **75.0** | — |
| | 5 | 7.5 | — | 10.0 | — | **82.5** |
| Group II | 1 | **95.0** | 5.0 | — | — | — |
| | 2 | 5.0 | **85.0** | — | 5.0 | 5.0 |
| | 3 | 2.5 | 10.0 | **82.5** | 5.0 | — |
| | 4 | — | — | 5.0 | **85.0** | 10.0 |
| | 5 | — | 25.0 | 10.0 | 5.0 | **60.0** |
| Group III | 1 | **55.0** | 25.0 | — | 15.0 | 5.0 |
| | 2 | 10.0 | **60.0** | 10.0 | 15.0 | 5.0 |
| | 3 | 10.0 | — | **70.0** | 10.0 | 10.0 |
| | 4 | 15.0 | 5.0 | 15.0 | **55.0** | 10.0 |
| | 5 | 5.0 | 5.0 | 15.0 | 10.0 | **65.0** |

Table 7: Experimental results.

| | Percentage of correct answers | | | | | Duration of a trial (second) |
|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | Ave./Std. |
| Group1 | 97.5 | 92.5 | 85.0 | 95.0 | 92.5 | 10.7/2.9 |
| Group2 | 92.5 | 100 | 77.5 | 97.5 | 75.0 | 13.4/4.0 |
| Group3 | 62.5 | 77.5 | 80.0 | 72.5 | 95.0 | 20.6/10.7 |

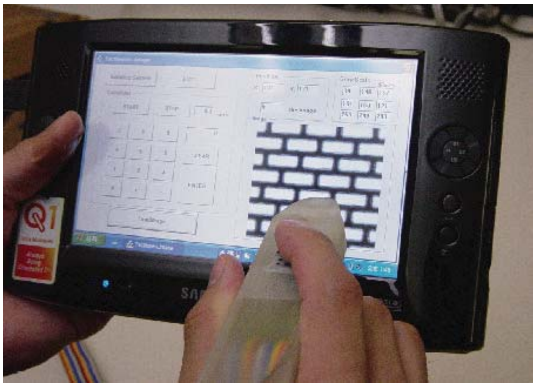


Figure 13: Tactile image display on a touchscreen.

texture display performance of the Ubi-Pen, the image samples from Section 5 were reused. One of five images from one of the groups was displayed on the screen, but hidden from the participant. Instead, the visual representation was of a blank square the same size as the image. When a user rubs against this square, the gray values from the image are presented to the tactile display on the Ubi-Pen. The experimental results are shown in Table 7 and these data verify that the Ubi-Pen and image display scheme are effective. This scheme can be applied to educational programs for children or interactive drawing software. In the future, this kind of technology could be the basis of a virtual interactive shopping mall.

### 6.2. *Medical applications*

One possible application of the combination of force and tactile feedback is a palpation medical simulator. Palpation is a kind of diagnosis based on pressure and pressure distribution. Therefore, when we develop a haptic palpation simulator, both force and tactile display interface are required. Kim et al. [27] proposed a palpation simulator based on this structure. However, their tactile display was somewhat cumbersome. The use of our tactile display or the Ubi-Pen might enhance the usability of this system; and there have been many other studies for haptic medical simulators which required a compact tactile display for more realistic and effective skin sense feedback.

### 6.3. *Additional applications*

As tested in Section 4.1, one of the most practical uses of our compact tactile display is Braille display. In particular, it can realize a highly portable Braille display. However, we need to conduct more precise evaluations before construction such a system.

Finally, the tactile display module could be installed in new mobile communication devices as well as PDAs and mobile computers.

## 7. CONCLUSION

This paper presents the Ubi-Pen II, a pen-like haptic interface with a built-in compact tactile display and an impact module, as well as empirical studies on Braille, button, and texture display. Its performance is verified in a series of preliminary evaluations which indicate that it can satisfactorily represent tactile patterns and provide impact feedback. The compact tactile display can represent Braille patterns and the impact feedback provides an effective button pressing sense which can increase user confidence. Furthermore, we investigated its applicability to combined force and tactile feedback interfaces in a haptic device with a pen-like end effecter. Force feedback, tactile feedback, and impact feedback have been compared for texture display. Of these three, combining tactile feedback with force feedback showed enhanced performance. Finally, we evaluated the Ubi-Pen II's capacity to support touch screen operations by providing tactile cues when a user rubs an image displayed on a monitor.

Future work involves improving the performance and usability of the Ubi-Pen II. To make the interface a stand-alone system, a processor and power controller should be embedded into the pen. The future version will be an interactive wireless interface; and more psychophysical and physiological studies will be involved in the next experiment for the Braille and texture display.

## ACKNOWLEDGMENTS

This work was supported by the IT R&D program of MIC/IITA (2007-S032-01, Development of an Intelligent Service technology based on the Personal Life Log). The authors appreciate Ian Oakley's kind editing and HIMS Corporation's support for the Braille display experiments.

## REFERENCES

[1] G. C. Burdea, *Force and Touch Feedback for Virtual Reality*, Wiley-Interscience, New York, NY, USA, 1996.

[2] Y. Ikei and M. Shiratori, "Texture explorer: a tactile and force display for visual textures," in *Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS '02)*, pp. 327–334, Orlando, Fla, USA, March 2002.

[3] I. R. Summers and C. M. Chanter, "A broadband tactile array on the fingertip," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2118–2126, 2002.

[4] K.-U. Kyung, M. Ahn, D.-S. Kwon, and M. A. Srinivasan, "A compact planar distributed tactile display and effects of frequency on texture judgment," *Advanced Robotics*, vol. 20, no. 5, pp. 563–580, 2006.

[5] V. Hayward and M. Cruz-Hernandez, "Tactile display device using distributed lateral skin stretch," in *Proceedings of the 8th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (ASME IMECE '00)*, vol. DSC-69-2, pp. 1309–1314, Orlando, Fla, USA, 2000.

[6] J. Luk, J. Pasquero, S. Little, K. MacLean, V. Lévesque, and V. Hayward, "A role for haptics in mobile interaction: initial design using a handheld tactile display prototype," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '06)*, vol. 1, pp. 171–180, Montreal, QC, USA, April 2006.

[7] M. Konyo, S. Tadokoro, and T. Takamori, "Artificial tactile feel display using soft gel actuators," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA '00)*, vol. 4, pp. 3416–3421, San Francisco, Calif, USA, April 2000.

[8] C. J. Poletto and C. Van Doren, "A high voltage stimulator for small electrode electrocutaneous stimulation," in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 6, pp. 2415–2418, Chicago, Ill, USA, October 1997.

[9] H. Kajimoto, N. Kawakami, T. Maeda, and S. Tachi, "Tactile feeling display using functional electrical stimulation," in *Proceedings of the 9th International Conference on Artificial Reality and Telexistence (ICAT '99)*, pp. 107–114, Tokyo, Japan, December 1999.

[10] D. A. Kontarinis and R. D. Howe, "Tactile display of vibratory information in teleoperation and virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 4, no. 4, pp. 387–402, 1995.

[11] M. Akamatsu and I. S. MacKenzie, "Movement characteristics using a mouse with tactile and force feedback," *International Journal of Human Computer Studies*, vol. 45, no. 4, pp. 483–493, 1996.

[12] P. Kammermeier, A. Kron, J. Hoogen, and G. Schmidt, "Display of holistic haptic sensations by combined tactile and kinesthetic feedback," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 1, pp. 1–15, 2004.

[13] R. J. Webster, T. E. Murphy, L. N. Verner, and A. M. Okamura, "A novel two-dimensional tactile slip display: design, kinematics and perceptual experiments," *ACM Transactions on Applied Perception*, vol. 2, no. 2, pp. 150–165, 2005.

[14] K.-U. Kyung, D.-S. Kwon, and G.-H. Yang, "A novel interactive mouse system for holistic haptic display in a human-computer interface," *International Journal of Human-Computer Interaction*, vol. 20, no. 3, pp. 247–270, 2006.

[15] D. Allerkamp, G. Böttcher, F.-E. Wolter, A. C. Brady, J. Qu, and I. R. Summers, "A vibrotactile approach to tactile rendering," *The Visual Computer*, vol. 23, no. 2, pp. 97–108, 2007.

[16] R. S. Johansson and A. B. Vallbo, "Tactile sensibility in the human hand: relative and absolute densities of four types of mechanoreceptive units in glabrous skin," *Journal of Physiology*, vol. 286, pp. 283–300, 1979.

[17] K. O. Johnson and J. R. Phillips, "Tactile spatial resolution. I. Two-point discrimination, gap detection, grating resolution, and letter recognition," *Journal of Neurophysiology*, vol. 46, no. 6, pp. 1177–1192, 1981.

[18] S. J. Bolanowski Jr., G. A. Gescheider, R. T. Verrillo, and C. M. Checkosky, "Four channels mediate the mechanical aspects of touch," *Journal of the Acoustical Society of America*, vol. 84, no. 5, pp. 1680–1694, 1988.

[19] R. T. Verrillo, A. J. Fraoli, and R. L. Smith, "Sensation magnitude of vibrotactile stimuli," *Perception and Psychophysics*, vol. 7, pp. 366–372, 1969.

[20] K.-U. Kyung and J.-Y. Lee, "Design and applications of a pen-like haptic interface with texture and vibrotactile display," to appear in *IEEE Computer Graphics and Applications*.

[21] M. A. Srinivasan, "Surface deflection of primate fingertip under line load," *Journal of Biomechanics*, vol. 22, no. 4, pp. 343–349, 1989.

[22] J. C. Lee, P. H. Dietz, D. Leigh, W. S. Yerazunis, and S. E. Hudson, "Haptic pen: a tactile feedback stylus for touch screens," in *Proceedings of the Annual ACM Symposium on User Interface Softaware and Technology (UIST '04)*, pp. 291–294, Santa Fe, NM, USA, October 2004.

[23] J. C. Craig, "Vibrotactile pattern perception: extraordinary observers," *Science*, vol. 196, no. 4288, pp. 450–452, 1977.

[24] I. Darian-Smith and P. Kenins, "Innervation density of mechanoreceptive fibres supplying glabrous skin of the monkey's index finger," *Journal of Physiology*, vol. 309, pp. 147–155, 1980.

[25] Immersion Corporation, "TouchSense technology for the touch screen interface: adding tactile feedback to touch screen applications," 2006.

[26] T. H. Massie and J. K. Salisbury, "PHANTOM haptic interface: a device for probing virtual objects," in *Proceedings of the ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, vol. 55-1, pp. 295–299, Chicago, Ill, USA, November 1994.

[27] S.-Y. Kim, K.-U. Kyung, J. Park, and D.-S. Kwon, "Real-time area-based haptic rendering and the augmented tactile display device for a palpation simulator," *Advanced Robotics*, vol. 21, no. 9, pp. 961–981, 2007.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 743690, 9 pages
doi:10.1155/2008/743690

*Research Article*

# Mathematical Modeling of Cytotoxic Lymphocyte-Mediated Immune Response to Hepatitis B Virus Infection

**Changjiang Long,[1, 2] Huan Qi,[2] and Sheng-Hu Huang[3, 4]**

[1] *School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[2] *Institute of Systems Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[3] *The Saban Research Institute, Childrens Hospital Los Angeles, Los Angeles, CA 90027, USA*
[4] *University of Southern California University Hospital, Los Angeles, CA 90027, USA*

Correspondence should be addressed to Huan Qi, qihuan@hust.edu.cn

Received 1 September 2007; Accepted 23 December 2007

Recommended by Daniel Howard

Nowak's model of the human immunodeficiency virus (HIV) infection has been extensively and successfully used to simulate the interaction between HIV and cytotoxic lymphocyte- (CTL-) mediated immune response. However, this model is not available for hepatitis B virus (HBV) infection. As the enhanced recruitment of virus-specific CTLs into the liver has been an important novel concept in the pathogenesis of hepatitis B, we develop a specific mathematical model analyzing the relationship between HBV and the CTL-mediated immune response, and the indicator of the liver cell damage, alanine aminotransferase (ALT). The stability condition of the complete recovery equilibrium point at which HBV will be eliminated entirely from the body is discussed. A different set of parameters is used in the simulation, and the results show that the model can interpret the wide variety of clinical manifestations of HBV infection. The model suggests that a rapid and vigorous CTL response is required for resolution of HBV infection.

## 1. INTRODUCTION

Infection with the hepatitis B virus (HBV) is a major health problem, which can lead to cirrhosis and primary hepatocellular carcinoma (HCC). More than 2 billion people alive today have been infected by HBV. The population of HBV carrier is about 400 million, of whom 75% are located in Asia. Accordingly, HBV causes approximately 1 million deaths each year worldwide. In China alone, nearly 15 million new infections occur annually, more than 30 million people are chronically infected, and more than 350 thousand of them die each year from cirrhosis and HCC.

In order to find an efficient way to prevent and treat the infection, it is of great importance to understand the immunopathogenesis of HBV. Although molecular techniques have provided fundamental insight into the fine detail of the interaction between HBV and immune system, many biologically important questions are not primarily concerned with the molecular mechanisms of immune recognition but with the population dynamics of the immune response. Mathematical models are always needed to answer these questions. Studies on humans infected with HIV-1 and macaques infected with simian immunodeficiency virus (SIV) show that cytotoxic lymphocytes (CTLs) are critical in controlling virus replication [1]. Virus mutants of human immunodeficiency virus (HIV) and SIV are able to escape the dominant CTL response and become the major replicating strains in vivo [1]. In contrast to HIV and many other viruses, cell culture systems that allow efficient in vitro infection and passaging of virus are not available for HBV, which is hepatotropic and noncytopathic. Recent studies on HBV pathogenesis in animal models demonstrated that the enhanced recruitment of virus-specific CTLs into the liver cells is critical for the pathogenesis of both HBV infection and hepatocellular carcinoma [2, 3]. The most common mathematical model in HIV infection is presented by Nowak to explain the dynamics of CTL-mediated host immune response to HIV and the pathogenesis of AIDS [4]. Mathematical models, which are not based on CTL-mediated host response, have been proposed for modeling HBV or HCV infection and evaluating the effectiveness of antiviral therapy [5–8]. Even the models describing the interactions between host immune response and virus were built to explain the mechanism of acute hepatitis [9–11]. However, these models fail to explain the various outcomes of HBV infection [12]. A new mathematical model for dissecting the role of CTL in HBV diseases is needed for the following reasons.

(1) Being different kinds of viruses, HIV is cytopathic virus and HBV is a noncytopathic virus [12], that is, cells infected by HBV will not be killed by virus directly, cellular function and lifespan of HBV-infected hepatocytes are almost the same as that of the uninfected cells in vitro [13]. The death rate of noncytopathic virus-infected cells in the absence of immunity equals that of uninfected target cells [14]. The lifespan of HBV-infected cells varies greatly in vivo which is mainly due to the strength of the anti-HBV CTL response [15]. CTL will not only kill but also cure the infected hepatocytes by a nonlytic effector mechanism [16, 17]. The effect of CTL response should be considered in the model.

(2) The non-CTL models ignore the kinetics of hepatocyte replication. Actually, both uninfected and infected hepatocytes can replicate at the same rate [13]. Infected cells are generated not only from normal cells infected by HBV, but also from replication of its own [18, 19].

(3) In the non-CTL models, the equilibrium abundance of infected cells depends only on the immunological parameters [6]; but in fact the characteristic of HBV also has great influence on the result [12].

(4) The value of alanine aminotransferase (ALT) in the blood stream is generally taken to be an indicator of the liver cell damage [9]. ALT was not included in the non-CTL models.

## 2. MATERIALS AND METHODS

### 2.1. Mathematical models

The model contains six variables, that is, uninfected hepatocytes ($X$), infected hepatocytes ($Y$), total host hepatocytes ($N = X + Y$), free virus ($V$), a CTL response ($Z$), and ALT. The changes of population over time can be described by a system of differential equations.

The corresponding mathematics equations are

$$\begin{gathered}
\frac{dX}{dt} = F(N)X - d_1X - b_1XV + k_1YZ,\\
\frac{dY}{dt} = F(N)Y + b_1XV - d_1Y - (k_1 + k_2)YZ,\\
\frac{dV}{dt} = k_3Y - d_3V,\\
\frac{dZ}{dt} = (g_4 + k_4YZ)\left(1 - \frac{Z}{Z_{\max}}\right) - d_4Z,\\
\frac{d\mathrm{ALT}}{dt} = k_5(d_1X + d_1Y + k_2YZ) - d_5\mathrm{ALT},\\
F(N) = \frac{d_1}{N^3}, \quad N = X + Y,\\
X_0 = 1, \quad Y_0 = 0, \quad V_0 = 0, \quad Z_0 = \frac{g_4}{d_4},\\
\mathrm{ALT}_0 = k_5\frac{d_1}{d_5}d_1, \qquad N_0 = 1,
\end{gathered} \tag{1}$$

where $F(N)$ is the natural growth rate of hepatocytes. It is a monotonically decreasing function [20]. We took $F(N) = d_1/N^3$ [13]; $F(N) = d_1$ when $N = 1$ (without loss of generality, we take the cell and virus concentrations to be scaled such that in the uninfected system the total cell concentration is $N = 1$ [18, 19]).

Both uninfected and infected hepatocytes replicate at a rate $F(N)$ and die at a rate $d_1$, while uninfected ones are infected by virus at a rate $b_1XV$. The CTL response can activate two different pathways to eliminate a virus, either by killing the infected cells or by eliminating the virus from within the cell without killing it. Infected cells are assumed to be killed by the CTL response at a rate $k_2YZ$ and be cured by the CTL response at a rate $k_1YZ$. Infected cells produce free virus at a rate $k_3Y$ and free virus particles are removed at a rate $d_3V$. CTLs proliferation can be described by two terms $g_4$ and $k_4YZ$, where $g_4$ represents antigen-independent proliferation and $k_4YZ$ represents antigen-dependent proliferation. CTLs decay at a rate $d_4Z$. ALT is generated by the dead hepatocytes at a rate $k_5$ and decay at a rate $d_5$. All the variables and parameters of the above are nonnegative.

### 2.2. Equilibrium states analysis

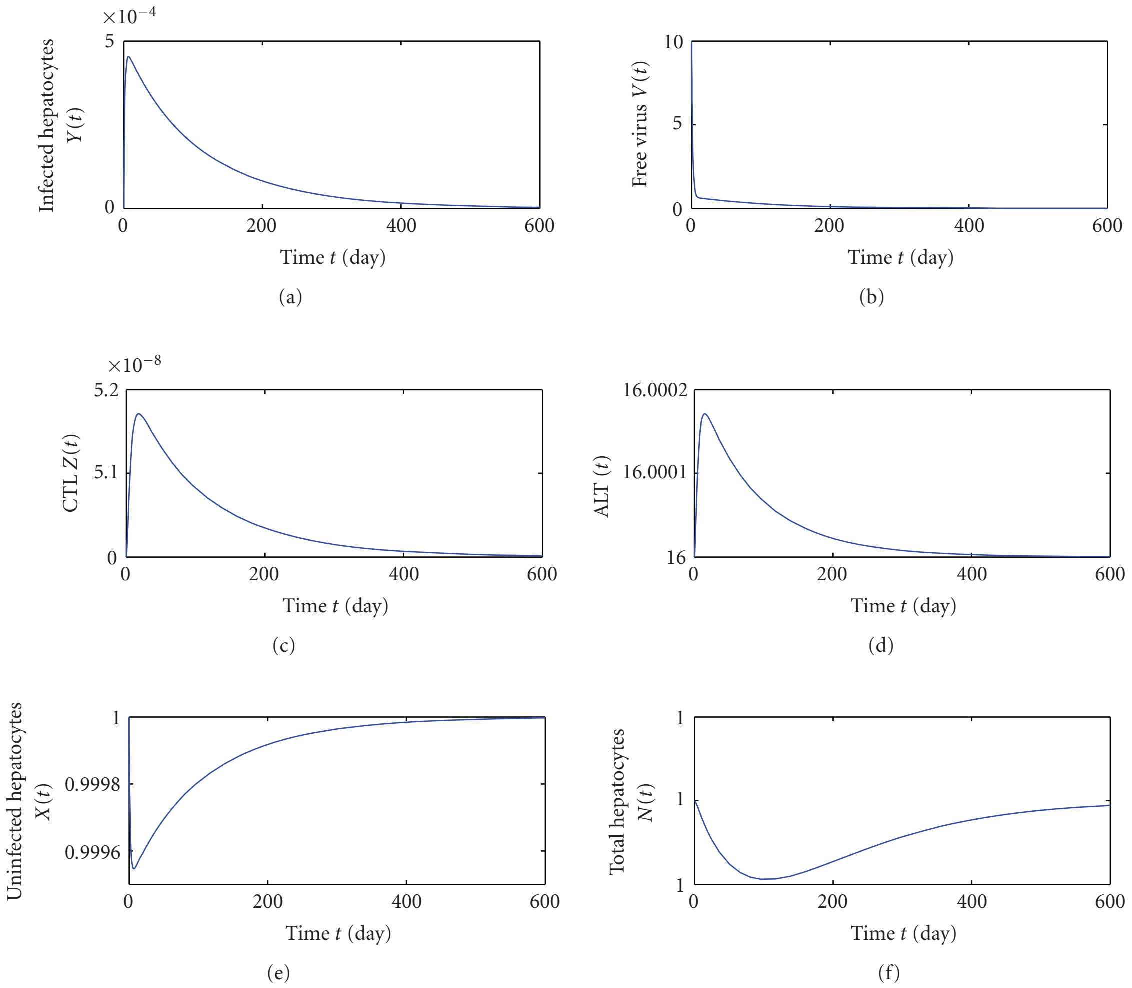
There are three possible steady states: Hepatocytes are not infected—the uninfected state, all the hepatocytes are infected—wholly infected state, and the coexisting state—the uninfected and the infected hepatocytes coexist. As these states are too complex to analyze, we only discuss the linear ability of the most concerned state—uninfected state.

As the equation of $Z$ is a logistic model, it will make the expression of the analysis result very complex. To get a meaning result easy for comparing with the clinical conclusion, we ignore the limitation of $Z_{\max}$ as we only discuss small disturbance to the initial state (the maximum number of specific T-cells can be $10^6$ times of its initial number [20]).

As ALT is just an indicator of hepatic injury and do not influence other variables, it is not included in the analysis progress for simplicity.

Then the equations of the system can be rewritten as

$$\begin{gathered}
\frac{dX}{dt} = F(N)X - d_1X - b_1XV + k_1YZ,\\
\frac{dY}{dt} = F(N)Y + b_1XV - d_1Y - (k_1 + k_2)YZ,\\
\frac{dV}{dt} = k_3Y - d_3V,\\
\frac{dZ}{dt} = g_4 + k_4YZ - d_4Z,\\
F(N) = \frac{d_1}{N^3}, \quad N = X + Y.
\end{gathered} \tag{2}$$



(a) (b) (c) (d) (e) (f)

Figure 1: Acute hepatitis.

For the uninfected state ($X = 1$, $Y = 0$, $V = 0$, $Z = g_4/d_4$), as it is an equilibrium state, it should satisfy $dX/dt = dY/dt = dV/dt = dZ/dt = 0$, so get the coordinates $(1, 0, 0, g_4/d_4)$. The correspondence Jacobian matrix is

$$\begin{bmatrix} F'(N) & F'(N) + \dfrac{k_1 g_4}{d_4} & -b_1 & 0 \\ 0 & \dfrac{-(k_1+k_2)g_4}{d_4} & b_1 & 0 \\ 0 & k_3 & -d_3 & 0 \\ 0 & \dfrac{k_4 g_4}{d_4} & 0 & -d_4 \end{bmatrix}. \tag{3}$$

The characteristic equation is

$$(\lambda - F'(N))(\lambda + d_4)(\lambda^2 + (d_3 + \Delta)\lambda + d_3\Delta - b_1 k_3) = 0, \tag{4}$$

where

$$\Delta = (k_1 + k_2)\frac{g_4}{d_4},$$

$$\frac{\partial N}{\partial X} = \frac{\partial (X+Y)}{\partial X} = 1; \qquad \frac{\partial N}{\partial Y} = \frac{\partial (X+Y)}{\partial Y} = 1,$$

$$\frac{\partial F(N)}{\partial X} = \frac{\partial F(N)}{\partial N}\frac{\partial N}{\partial X} = F'(N), \tag{5}$$

$$\frac{\partial F(N)}{\partial Y} = \frac{\partial F(N)}{\partial N}\frac{\partial N}{\partial Y} = F'(N),$$

$$F'(N) = \frac{\partial F(N)}{\partial N} = \frac{\partial (d_1/N^3)}{\partial N} = \frac{-3d_1}{N^4} = -3d_1.$$

Since $-F'(N) = 3d_1 > 0$, $d_4 > 0$, $d_3 > 0$, $\Delta > 0$, $d_3 + \Delta > 0$, according to the Routh-Hurwitz criterion [21], if $d_3\Delta - b_1 k_3 > 0$, that is,

$$(k_1 + k_2)\frac{g_4}{d_4} > b_1\frac{k_3}{d_3}, \tag{6}$$

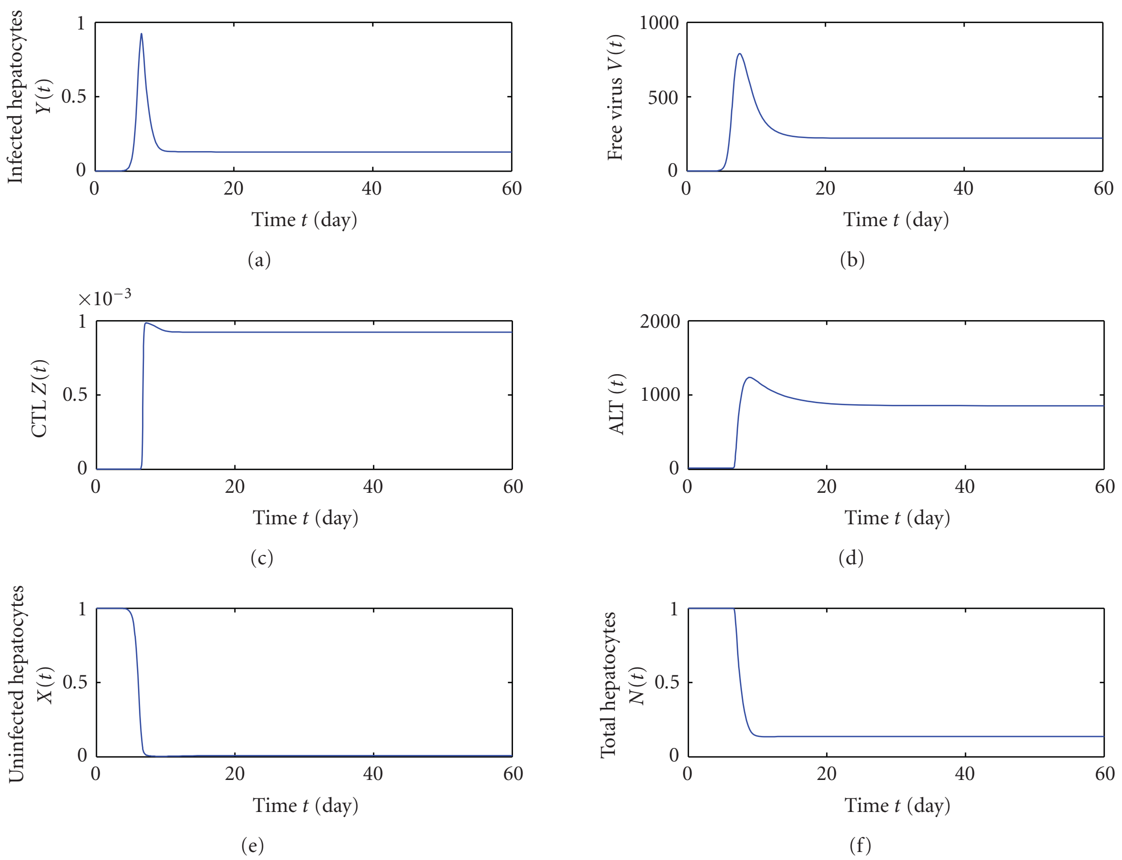


Figure 2: Fulminant hepatitis.

there will be linear stability with respect to perturbations in $(1, 0, 0, g_4/d_4)$. The left-hand side of the equation represents the ability of the immune system, $k_1$ represents the nonlytic ability of CTLs and $k_2$ represents the lytic ability of CTLs. Also, $g_4/d_4$ is the initial value of CTLs. The right-hand side of the equation represents the ability of HBV, $b_1$ represents the infective ability of HBV, $k_3$ represents the multiplication ability of HBV, and $d_3$ represents the death rate of HBV. If the left part is bigger than the right part, it means that the immune system is strong enough to eliminate the infection otherwise HBV can invade the body and exist for a long time.

### 2.3. Simulation

By combining the various results derived in the previous section we can deduce an appropriate parameter set for the simulation of the model.

The parameters (1 time unit = 1 day) are set up as follows: $d_1 = 0.002$ (average life of hepatocyte is about 500 d [13]); $d_3 = 0.58$ (estimated average half-life of free virions is about 1.2 d [22]); $d_4 = 0.2$ (the mean life of CTL is 4–6 d [20]); $d_5 = 0.25$ (average half-life of ALT is between 0.5–5 [13]); $Z_{\max} = 0.001$ there should be no more than $10^8$ HBV-specific CTL in the entire body and there are approximately $10^{11}$ infected hepatocytes in the human liver [23], as we took $N = 1$, so $Z_{\max} = 10^8/10^{11}$); $\text{ALT}_0 = 16$ (The normal range for ALT is between 0–40).

## 3. RESULTS

### 3.1. Acute hepatitis

HBV can cause acute hepatitis, resulting in short-term inflammation of the liver before the immune system is able to remove the virus from the body. In acutely infected patients who successfully control the virus, the immune response that the patients produce against the viral proteins is polyclonal, multispecific; and the virus is eliminated from the blood and liver entirely. If the maximum damage and the maximum concentration of free virus are low, the disease may come and go without any symptoms, otherwise severe clinical symptoms will be observed.

The parameters during the simulation of acute hepatitis are set up as follows: $d_1 = 0.002$; $b_1 = 3E - 5$; $k_1 = 1E6$; $k_2 = 1E3$; $b_3 = 0.1$; $k_3 = 800$; $d_3 = 0.58$; $g_4 = 1E - 8$; $k_4 = 16$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 10$; $Z_0 = 5E - 8$; $\text{ALT}_0 = 16$. As the cellular immune response is vigorous and $(k_1 + k_2)(g_4/d_4) = 0.0500 > b_1(k_3/d_3) = 0.0414$, so the immune system is strong enough to eliminate all the infected cells and virus. The number of virus is at its peak at the beginning and then decrease. HBV can only infect a small number of cells. The number of infected cells $(4.53 \times 10^{-4})$ peaked at 7 days after infection. As the number is so small that the level of serum ALT and the number of total hepatocytes almost

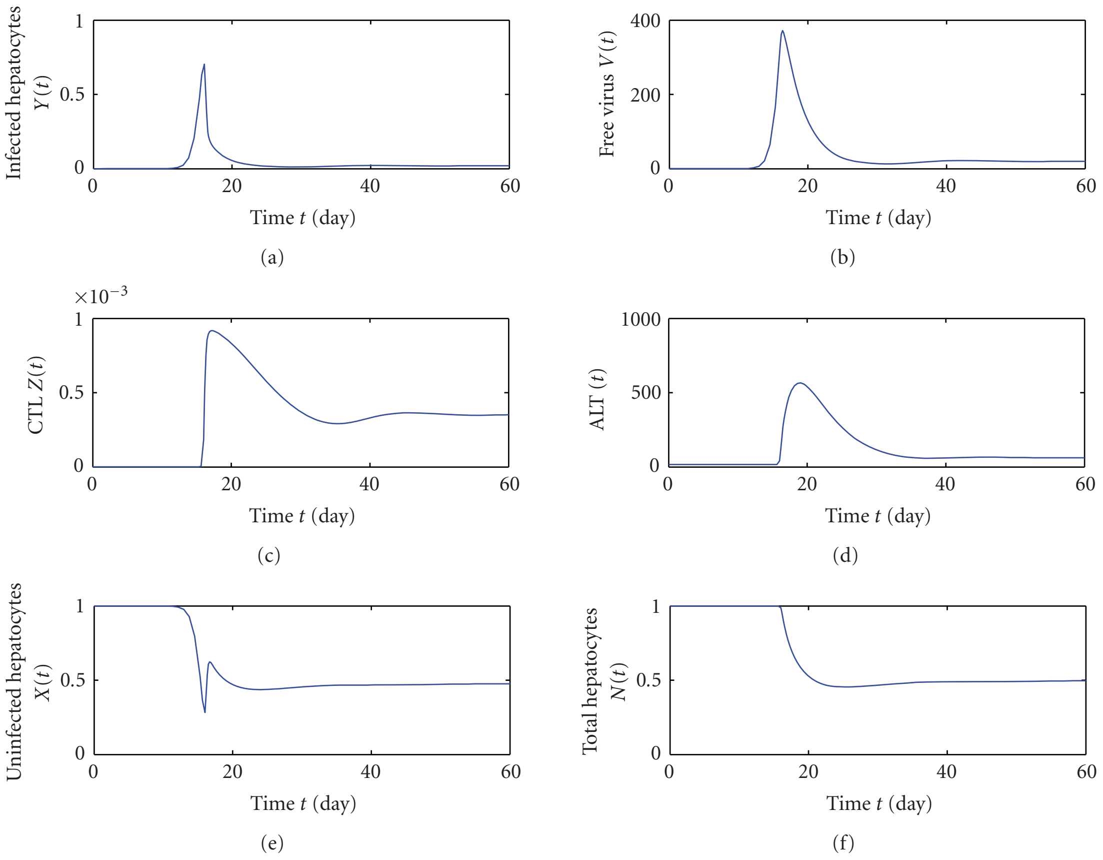


Figure 3: Acute–turn-chronic hepatitis.

stay steady when the infected cells are eliminated by the CTLs, viral clearance occur rapidly and efficiently with little evidence of liver disease. Simulation results are shown in Figure 1.

When the $(k_1 + k_2)(g_4/d_4) - b_1(k_3/d_3) < 0$, the cellular immune response is not able to eliminate all the infected hepatocytes, so the virus will be persistent. In this case, if the immune response is strong and many infected cells are killed, it will be chronic hepatitis B. If the immune response is very weak, there will be no symptom. If the virus has a strong infectious capability and the immune response is vigorous but not enough to resolve the infection, it will be fulminant hepatitis.

### 3.2. Fulminant hepatitis

The parameters during the simulation of fulminant hepatitis are set up as follows: $d_1 = 0.002$; $b_1 = 0.01$; $k_1 = 100$; $k_2 = 900$; $b_3 = 0.1$; $k_3 = 1000$; $d_3 = 0.58$; $g_4 = 1E-10$; $k_4 = 20$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 1E-5$; $Z_0 = 5E-8$; $\mathrm{ALT}_0 = 16$.

The mortality of fulminant hepatitis is up to 60% ~ 90%. In this case, the virus rapidly replicates and infects every hepatocyte in the liver. Most infected hepatocytes are destructed by the CTLs, resulting in severe liver dysfunction. Simulation results of fulminant hepatitis are shown in Figure 2.

The initiation value of $V_0$ is set as $10^{-5}(10^6/10^{11})$, as we assume transfusing 100 mL of blood from an inactive carrier whose serum HBV DNA level was about $10^4$ copies/mL would deliver $1 \times 10^6(100 \times 10^4)$ particles of HBV into the body.

As shown in Figure 2, hepatitis would outbreak sharply 7 days after the virus entered the body of a host. As the virus has a strong infectious capability $(b_1 = 0.01)$ and replicates rapidly $(k_3 = 1000)$, 90% of the hepatocytes in the liver will get infected within two days. The total number of HBV particles peaked at $794 \times 10^{11}$ (equals serum HBV DNA level $= 2.6 \times 10^{10}$ copies/mL). As the cellular immune response is rapidly elicited $(k_4 = 20)$, CTLs soon arrive the maximum value. The cytopathic effect of CTLs is more powerful than its noncytopathic effect $[(k_2 = 900) > (k_1 = 100)]$. Most infected cells are killed by CTL directly and this would lead to serious liver necrosis, and the level of ALT starts to rise sharply, it peaks at 1223 at 7 day.

### 3.3. Acute–turn-chronic hepatitis

The parameters during the simulation of acute–turn-chronic hepatitis are set up as follows: $d_1 = 0.002$; $b_1 = 0.005$; $k_1 = 6000$; $k_2 = 1000$; $b_3 = 0.1$; $k_3 = 600$; $d_3 = 0.58$; $g_4 = 1E-10$; $k_4 = 16$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 1E-7$; $Z_0 = 5E-8$; $\mathrm{ALT}_0 = 16$.

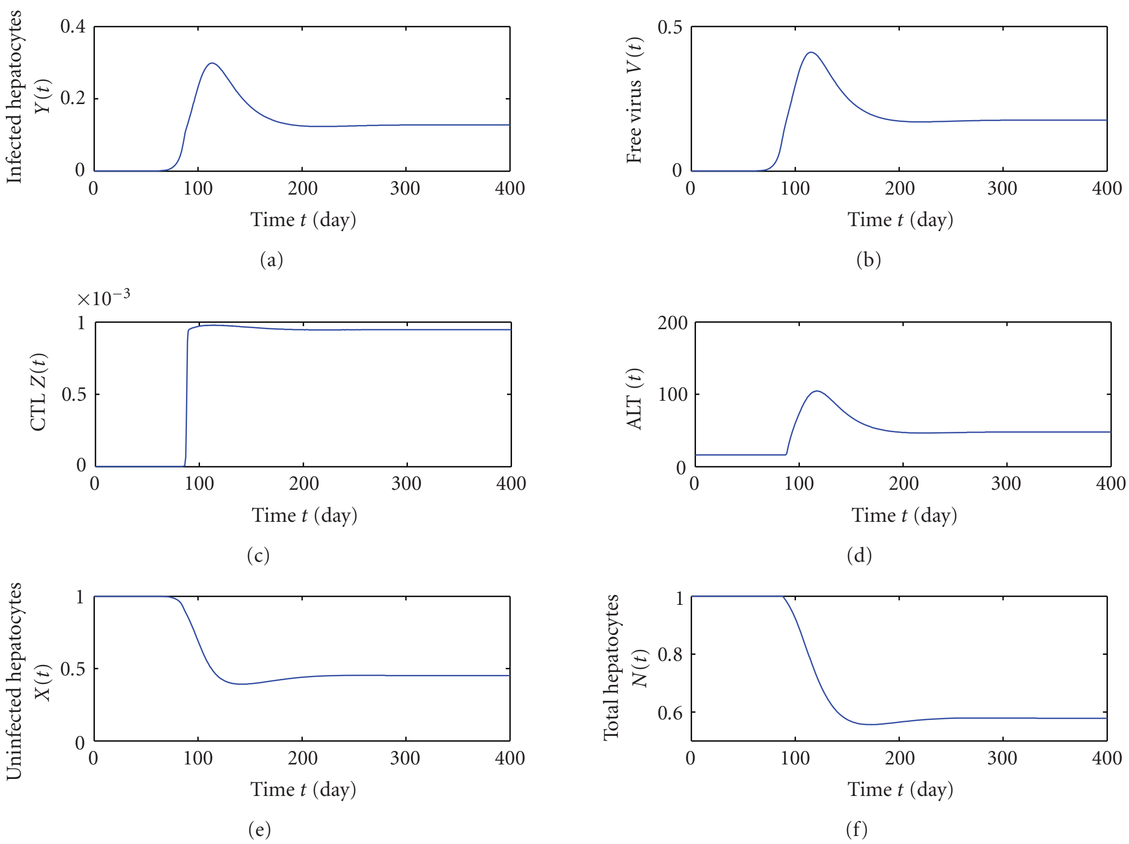


Figure 4: Chronic hepatitis.

HBV infection can become a chronic infection when the immune system cannot fight off the virus within six months after infection. It will establish a chronic, lifelong infection in the liver, and will have an enormously increased risk of developing liver cancer. It is well known that the T cell response is much less vigorous in chronically infected patients than it is during acute infection. Simulation results of acute–turn-chronic hepatitis are shown in Figure 3.

The initiation value of $V_0$ is set as $10^{-7}(10^4/10^{11})$, as we assume transfusing 1 mL of blood from an inactive carrier whose serum HBV DNA level was about $10^4$ copies/mL would deliver $1 \times 10^4 = 10^4$ particles of HBV into the body.

The incubation period of hepatitis B caused by the virus or blood transfusions is about 14 to 180 days. As shown in Figure 3, hepatitis would outbreak 16 days after the virus entered the body of a host. Eighty percent of the hepatocytes in the liver would get infected within 16 days, total number of HBV peaked at 371 (equals serum HBV DNA level $= 1.2 \times 10^{10}$ copies/mL). As the cytopathic effect of CTLs is so powerful ($k_2 = 1000$), many infected cells were killed by CTLs, and the level of ALT peaked to 569 at 19 day.

### 3.4. *Chronic hepatitis without acute phase*

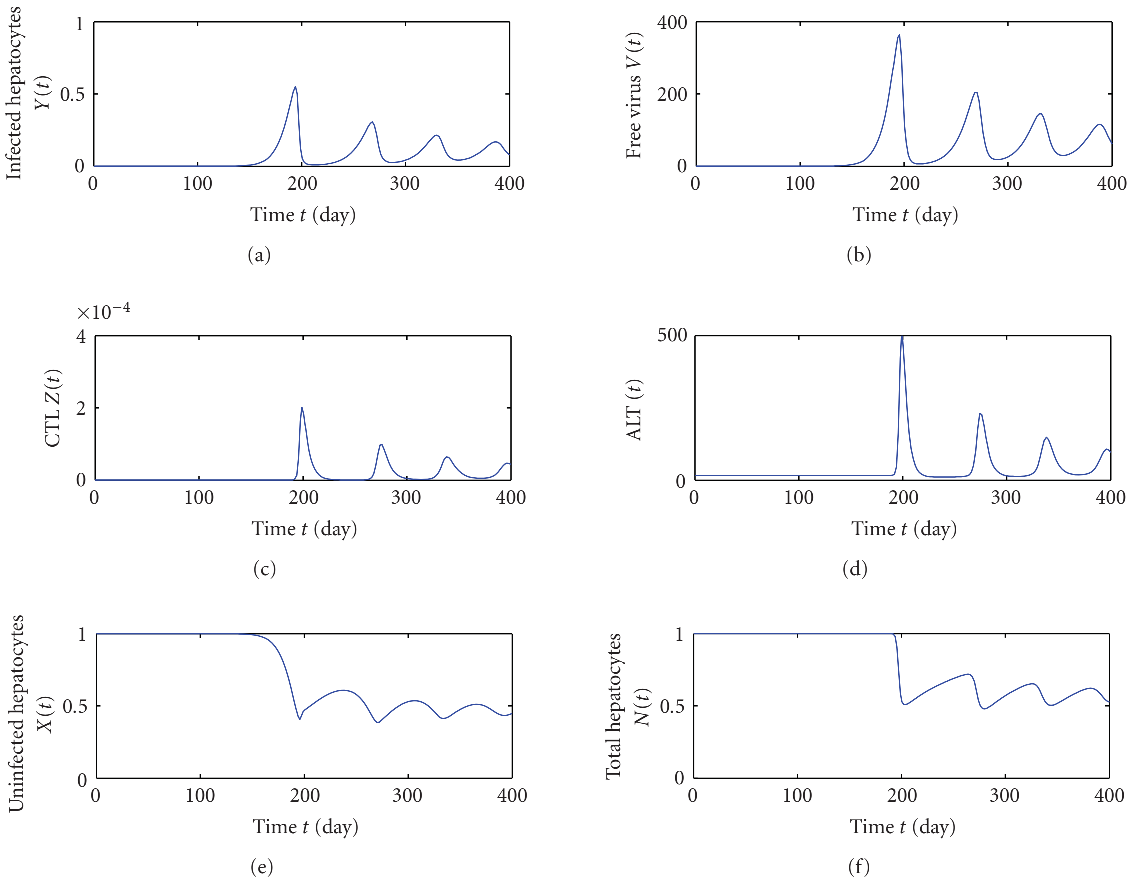
The parameters during the simulation of chronic hepatitis without acute phase are set up as follows: $d_1 = 0.002$; $b_1 = 0.2$; $k_1 = 100$; $k_2 = 40$; $b_3 = 0.1$; $k_3 = 0.8$; $d_3 = 0.58$; $g_4 = 1E-10$; $k_4 = 30$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $g_6 = 1E-10$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 1E-8$; $Z_0 = 5E-8$; $\text{ALT}_0 = 16$.

Many chronically infected people show little or no clinical signs. The HBV-specific immune response is too weak to eliminate HBV from all infected hepatocytes, but it is strong enough to continuously destroy HBV-infected hepatocytes, maybe resulting in progressive tissue damage and even cancer. Simulation results of chronic hepatitis without acute phase are shown in Figure 4.

The incubation of the Hepatitis B is about 42 to180 days, average 180 days. The initiation value of $V_0$ is set as $10^{-8}(10^3/10^{11})$, as we assume that a hypodermic needle carrying HBV-contaminated blood which circulates through the body will spread $10^3$ virions. As shown in Figure 4, hepatitis would outbreak 51 days after the virus entered the body. Maximally 35% of the hepatocytes in the liver would get infected, total number of HBV peaked at 0.56 (equals serum HBV DNA level $= 1.8 \times 10^7$ copies/mL). The level of ALT reaches its peak value (104) at 117 day. The system will arrive its steady state at 175 day, 15% of the hepatocytes in the liver are infected cells, and total number of HBV is 0.24 (equals serum HBV DNA level $= 8 \times 10^6$ copies/mL) and the ALT level is 48.

### 3.5. *Recurring hepatitis*

The parameters during the simulation of recurring hepatitis are set up as follows: $d_1 = 0.002$; $b_1 = 0.0002$; $k_1 = 1000$;



(a) (b) (c) (d) (e) (f)

Figure 5: Recurring hepatitis.

$k_2 = 2500$; $b_3 = 0.1$; $k_3 = 400$; $d_3 = 0.58$; $g_4 = 1E - 10$; $k_4 = 2.1$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 1E - 6$; $Z_0 = 5E - 8$; $\mathrm{ALT}_0 = 16$.

When the viral concentration is at its lowest, the patient may be diagnosed as complete recovery; but the virus never completely disappears and an apparent reinfection will soon appear. This recurrence will last for years. The simulation results of recurring hepatitis are shown in Figure 5.

### *3.6. Asymptomatic chronic hepatitis*

The parameters during the simulation of asymptomatic chronic hepatitis are set up as follows: $d_1 = 0.002$; $b_1 = 0.0001$; $k_1 = 1$; $k_2 = 4$; $b_3 = 0.1$; $k_3 = 100$; $d_3 = 0.58$; $g_4 = 1E - 10$; $k_4 = 10$; $d_4 = 0.2$; $k_5 = 2000$; $d_5 = 0.25$; $X_0 = 1$; $Y_0 = 0$; $V_0 = 1E - 7$; $Z_0 = 5E - 8$; $\mathrm{ALT}_0 = 16$.

Vertical transmission of HBV results in milder hepatitis in patients with no symptoms. The virus establishes itself in this immunologically immature population and is tolerated so that there will be no adequate immune response. Neonatal tolerance is probably responsible for both the lack of an antiviral immune response and the viral persistence after mother–infant transmission. This is the most common antecedent of persistent HBV infection worldwide. The simulation results of asymptomatic chronic hepatitis are shown in Figure 6.

## 4. DISCUSSION

The diversity of clinical syndromes and disease manifestations associatedwith HBV infection strongly suggests that the clinical outcome of this infection is determined by host-virus interactions, especially the quality and vigor of the antiviral immune response produced by the infected host. Most perinatal HBV infections become persistent, presumably due to a suboptimalcellular immune response that destroys some of the infected hepatocytes and does not purge the virus from the remaining infected hepatocytes. It thereby permits the persisting virus to trigger a chronic indolent necroinflammatory liver disease that sets the stage for development of HCC. In contrast, most of the adult onset HBV infections resolve, presumably due to the polyclonal, multispecific cellular immune response that the patients produce against the viral proteins [24].

The qualitative analysis and simulation results suggest the following pattern.

If the cellular immune response is vigorous and satisfied $(k_1 + k_2)g_4/d_4 > b_1k_3/d_3$, the immune system is strong enough to eliminate the infection. Otherwise, chronic hepatitis appears.

If the virus with strong infectious capability ($b_1$ is large) replicates rapidly ($k_3$ is large), most hepatocytes in the liver

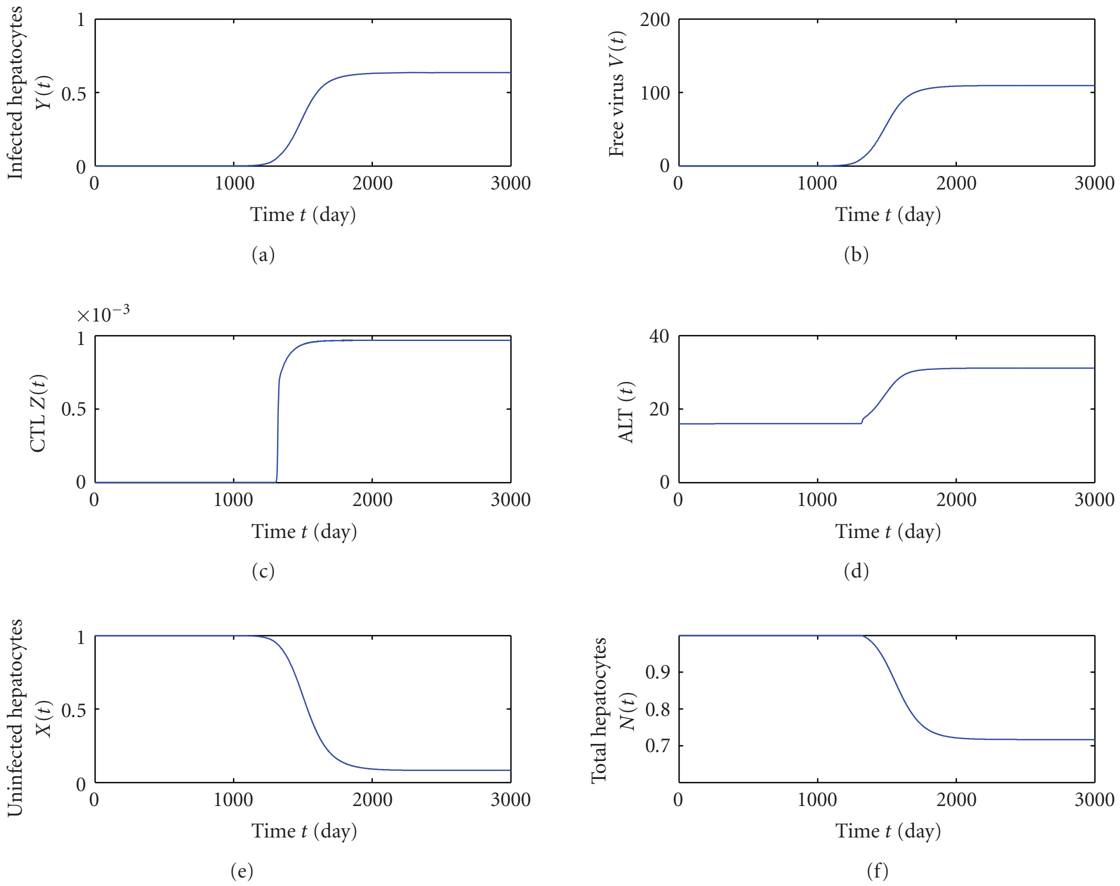


Figure 6: Asymptomatic hepatitis.

get infected, resulting in massive liver necrosis due to the strong CTL response. The outcome will be fulminant hepatitis.

If the virus with weak infectious capability replicates slowly, the CTL response to HBV is rapid ($k_4$ is large) and vigorous ($k_1 + k_2$ is large) enough to eliminate the virus from the blood and liver entirely. The outcome will be acute hepatitis. If the maximum damage and the maximum concentration of free virus are low, the disease may come and go unnoticed, otherwise severe clinical symptoms will be observed.

If the immune system defends against HBV with a weak killing ability ($k_1 + k_2$ is small) and weak CTL level ($k_4$ is small), the infected cells cannot be cleared out entirely. The outcome will be chronic hepatitis with little or no clinical signs.

This model is able to account for the different outcomes of HBV infection. However, this model can be further improved to explain why many patients with acute hepatitis, whose hepatocytes are almost all infected, can recover, and why many patients with chronic hepatitis B can get rid of HBV when getting older. For the future studies, the model will be applied to fit clinical data for the evaluation of immune states and virus characteristic, thus providing information about the potency of antiviral therapies and guiding the development of optimal drug dosages and regimens.

## ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China Grant no. 60774036 (H.Q).

## REFERENCES

[1] P. A. Morel, “Mathematical modeling of immunological reactions,” *Frontiers in Bioscience*, vol. 3, no. 3, pp. 338–337, 1998.

[2] S. M. Ciupe, R. M. Ribeiro, P. W. Nelson, G. Dusheiko, and A. S. Perelson, “The role of cells refractory to productive infection in acute hepatitis B viral dynamics,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 12, pp. 5050–5055, 2007.

[3] B. Xuli and D. Zhongping, “Advance in viral dynamics of hepatitis B virus infection,” *Foreign Medical Science (Section of Virology)*, vol. 11, no. 2, pp. 36–40, 2004.

[4] M. A. Nowak and C. R. M. Bangham, “Population dynamics of immune responses to persistent viruses,” *Science*, vol. 272, no. 5258, pp. 74–79, 1996.

[5] A. S. Perelson and R. M. Ribeiro, “Hepatitis B virus kinetics and mathematical modeling,” *Seminars in Liver Disease*, vol. 24, supplement 1, pp. 11–16, 2004.

[6] M. A. Nowak, S. Bonhoeffer, A. M. Hill, R. Boehme, H. C. Thomas, and H. Mcdade, “Viral dynamics in hepatitis B virus infection,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 9, pp. 4398–4402, 1996.

[7] R. J. H. Payne, M. A. Nowak, and B. S. Blumberg, "The dynamics of hepatitis B virus infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 13, pp. 6542–6546, 1996.

[8] P. Colombatto, L. Civitano, R. Bizzarri, et al., "A multiphase model of the dynamics of HBV infection in HBeAg-negative patients during pegylated interferon-$\alpha$2a, lamivudine and combination therapy," *Antiviral Therapy*, vol. 11, no. 2, pp. 197–212, 2006.

[9] S. M. Ciupe, R. M. Ribeiro, P. W. Nelson, and A. S. Perelson, "Modeling the mechanisms of acute hepatitis B virus infection," *Journal of Theoretical Biology*, vol. 247, no. 1, pp. 23–35, 2007.

[10] A. S. Perelson, E. Herrmann, F. Micol, and S. Zeuzem, "New kinetic models for the hepatitis C virus," *Hepatology*, vol. 42, no. 4, pp. 749–754, 2005.

[11] H. Dahari, M. Major, X. Zhang, et al., "Mathematical modeling of primary hepatitis C infection: noncytolytic clearance and early blockage of virion production," *Gastroenterology*, vol. 128, no. 4, pp. 1056–1066, 2005.

[12] Y. Ilan, "Immune down regulation leads to up regulation of an antiviral response: a lesson from the hepatitis B virus," *Microbes and Infection*, vol. 4, no. 13, pp. 1317–1326, 2002.

[13] L. Kangxian, *Hepatitis B Basic Biology and Clinical Science*, People's Medical Publishing House, Beijin, China, 2006.

[14] D. Wodarz, "Mathematical models of immune effector responses to viral infections: virus control versus the development of pathology," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 301–319, 2005.

[15] M. A. Nowak and M. M. Robert, *Virus Dynamics: Mathematical Principles of Immunology and Virology*, Oxford University Press, New York, NY, USA, 2000.

[16] A. Bertoletti, M. Maini, and R. Williams, "Role of hepatitis B virus specific cytotoxic T cells in liver damage and viral control," *Antiviral Research*, vol. 60, no. 2, pp. 61–66, 2003.

[17] L. G. Guidotti, "Pathogenesis of viral hepatitis," *Biological Regulators and Homeostatic Agents*, vol. 17, no. 2, pp. 115–119, 2003.

[18] R. J. H. Payne, M. A. Nowak, and B. S. Blumberg, "A cellular model to explain the pathogenesis of infection by the hepatitis B virus," *Mathematical Biosciences*, vol. 123, no. 1, pp. 25–58, 1994.

[19] R. J. H. Payne, M. A. Nowak, and B. S. Blumberg, "Analysis of a cellular model to account for the natural history of infection by the hepatitis B virus and its role in the development of primary hepatocellular carcinoma," *Journal of Theoretical Biology*, vol. 159, no. 2, pp. 215–240, 1992.

[20] Q. Sheng and D. Chanying, *Nonlinear Models in Immunity*, Shanghai Science & Technology Education Press, Shanghai, China, 1998.

[21] L. Yun, *Theory for Nonlinear Dynamic System of Modern Mathematics and Its Application*, Communications Press, Beijing, China, 1998.

[22] J. M. Murray, S. F. Wieland, R. H. Purcell, and F. V. Chisari, "Dynamics of hepatitis B virus clearance in chimpanzees," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 49, pp. 17780–17785, 2005.

[23] F. V. Chisari, "Viruses, immunity, and cancer: lessons from hepatitis B," *The American Journal of Pathology*, vol. 156, no. 4, pp. 1117–1132, 2000.

[24] M. Iannacone, G. Sitia, Z. M. Ruggeri, and L. G. Guidotti, "HBV pathogenesis in animal models: recent advances on the role of platelets," *Journal of Hepatology*, vol. 46, no. 4, pp. 719–726, 2007.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 375620, 7 pages
doi:10.1155/2008/375620

*Research Article*

# Infectomic Analysis of Gene Expression Profiles of Human Brain Microvascular Endothelial Cells Infected with *Cryptococcus neoformans*

**Ambrose Jong,[1] Chun-Hua Wu,[1] Wensheng Zhou,[2] Han-Min Chen,[3] and Sheng-He Huang[1]**

[1] *Divisions of Hematology-Oncology and Infectious Diseases, Saban Research Institute of Childrens Hospital Los Angeles, Keck School of Medicine, University of Southern California, Los Angeles, CA 90027, USA*
[2] *HRL Laboratories, LLC, Malibu, CA 90265, USA*
[3] *Department of Life Science, Fu Jen Catholic University, Hsinchuang, Taipei Country 24205, Taiwan*

Correspondence should be addressed to Ambrose Jong, ajong@chla.usc.edu

Received 12 August 2007; Accepted 17 December 2007

Recommended by Daniel Howard

In order to dissect the pathogenesis of *Cryptococcus neoformans* meningoencephalitis, a genomic survey of the changes in gene expression of human brain microvascular endothelial cells infected by *C. neoformans* was carried out in a time-course study. Principal component analysis (PCA) revealed significant fluctuations in the expression levels of different groups of genes during the pathogen-host interaction. Self-organizing map (SOM) analysis revealed that most genes were up- or downregulated 2 folds or more at least at one time point during the pathogen-host engagement. The microarray data were validated by Western blot analysis of a group of genes, including $\beta$-actin, Bcl-x, CD47, Bax, Bad, and Bcl-2. Hierarchical cluster profile showed that 61 out of 66 listed interferon genes were changed at least at one time point. Similarly, the active responses in expression of MHC genes were detected at all stages of the interaction. Taken together, our infectomic approaches suggest that the host cells significantly change the gene profiles and also actively participate in immunoregulations of the central nervous system (CNS) during *C. neoformans* infection.

## 1. INTRODUCTION

The major challenge posed by infectious diseases is to holistically and integratively understand the fundamental issues of how infectious agents and human hosts interact during microbial infection.Host-microbe interactions in the pathogenesis of infectious diseases are dynamic and complex processes [1] which result in changes in whole genome expression profiles of microbial pathogens and their hosts. A new discipline called infectomics [1] became established when the recently developed high-throughput omic approaches and computational tools were combined with the conventional approaches for the study of infectious diseases. Infectomes are detailed maps of microbial infections, and the availability of whole genomes of many living organisms paves the way for their holistic and integrative study. Infectomics can be defined as the study of infectomes, which are encoded by genomes of microbes and their hosts. Microarray has been a powerful tool to monitor infectomes in microorganisms and their host responses during microbial infection.

Infection by *Cryptococcus neoformans* has increased considerably over the past few years [2–4]. Dehydrated haploid yeast or basidiospore of *C. neoformans* is the usual form of inhalation [3, 5]. The organisms are likely to spread hematogeneously to extrapulmonary tissues and show a remarkable propensity in spreading to the brain and meninges, where life-threatening meningoencephalitis develops [2, 6, 7]. In order to cause meningoencephalitis, *C. neoformans* must penetrate the blood-brain barrier (BBB), which is a barrier between blood circulation and the brain parenchyma. BBB mainly consists of brain microvascular endothelial cells (BMECs), which are responsible for maintaining the biochemical homeostasis within the central nervous system (CNS) [8–10]. BMEC has been established as an in vitro cell culture model for dissecting the underlying mechanism(s) whereby *C. neoformans* crosses the BBB. We have recently demonstrated that *C. neoformans* are able to alter the cytoskeleton of human brain microvascular endothelial cells (HBMECs) [11]. We have also identified and characterized a *C. neoformans* capsule gene, *CPS1* [12]. This demonstrated that *CPS1* encodes hyaluronic acid synthase. The above information suggested that *C. neoformans* hyaluronic acid (HA) plays a role as an adhesion molecule during the yeast entry. It also suggested that host cell factors are required for *C. neoformans* HA-binding and the pathogen entry into HBMEC [12].

Like many other pathogens, *C. neoformans* may manipulate the host system to facilitate its invasion. The investigation of virulence of the pathogen *C. neoformans* and the study of the responses from HBMEC are equally important in the understanding of the complex invasion process. A more comprehensive knowledge of the interplay between the host and microbial pathogen at the levels of genome expression profiles is central to the understanding of the pathogenesis of infectious diseases. In order to dissect the pathogenesis of this disease, we have combined the infectomic approach with the in vitro model of the BBB to monitor gene expression profiles of HBMECs infected with *C. neoformans*. These studies provide global and useful information for building a comprehensive framework to interpret *C. neoformans* pathogenic processes.

## 2. MATERIALS AND METHODS

### 2.1. *Cultures of yeasts and human brain microvascular endothelial cells (HBMECs)*

*C. neoformans* strain B3501 was used for this study [11]. Yeast cells were grown aerobically at 30°C in the rich YPD broth containing 1% yeast extracts, 2% peptone, and 2% dextrose. Cells were harvested at early log phase, washed with PBS, and resuspended in Ham-F12/M199 (1:1; v:v) and 5% heat inactivated fetal bovine serum (FBS) (experimental medium).

The HBMEC culture was prepared as described previously [11, 13]. Briefly, the HBMEC cultures were maintained in RPMI 1640 medium including 10% FBS and 10% NuSerum (BD Biosciences, Bedford, MA, USA) at 37°C in a humid atmosphere of 5% $CO_2$ as described above. For the preparation of interactive cultures for microarray analysis, the HBMEC were grown in collagen-coated 24 well tissue culture plates (Costar Corp, Cambridge, MA, USA) until confluency. An inoculum of $10^6$ yeast cells in 1 mL experimental medium was added. *C. neoformans* B3501 was incubated with HBMEC at 37°C and harvested at 0, 4, 8, 12, 16, 20, 24 hours. One-tenth of cell pellets were saved for Western blots and the rest were subjected to the RNA extraction for making the probes.

### 2.2. *Preparation of the biotinylated cRNA for microarray*

Total RNA was preparedusing TRIZOL reagent (Invitrogen, Calif, USA), and subjectedto isolation of poly$(A)^+$ RNA using the Oligotex-dT30 mRNA purification kit (TaKaRa Shuzo Co., Kyoto, Japan) according to the manufacturer's instructions. The biotinylated cRNA probe was prepared using the RNA Transcript labeling kit (Enzo Biochem, Farmingdale, NY, USA) according to the manufacturer's instructions. The quality of the probe was first examined with 1% agarose gel, showing the bands between 0.5 ∼ 1.5 kb, peaking at around 0.75 kb. An aliquot of biotinylated cRNA probes was then examined by the Affymetirx test chips. Internal controls, G3PDH and $\beta$-actin, on the test chip were measured to evaluate the biotinylated probes. Only high-quality cRNA probes were used to hybridize with DNA microarray HU95A chips (12,809 gene spots in ∼1.64 $cm^2$ filters). After hybridization, the HP GeneArray scanner was used to analyze the patterns of gene expression.

### 2.3. *Microarray analysis*

Biotinylated cRNA probes were prepared from 10 $\mu$g of poly$(A)^+$ RNA. Hybridization and fluorescence detection were performed according to the manufacturer's instructions. Images were analyzed with GeneSpring, Genetrix softwares. Three clustering approaches were used in this study as follows. (a) Hierarchical clustering where the data points were organized in a phylogenetic tree in which the branch lengths represent the degree of similarity between the values. (b) Self-organizing maps (SOMs) that was a nonhierarchical clustering approach. Using this algorithm, gene expression data were transformed into vectors or coordinates in an $n$-dimensional space, where $n$ equals the number of variables or time points. (c) Principal-component analysis (PCA) that was used to obtain a simplified visualization of entire datasets.

### 2.4. *Western blot analysis*

Protein concentration was determined by Bio-Rad protein assay and equal amounts of protein were used from different time point samples. SDS̃-PAGE sample buffer (50 mM Tris-HCl pH 6.8, 10% $\beta$-mercaptoethanol, 2% SDS̃, 0.1% bromophenol blue, 10% glycerol) was added to the samples and they were boiled in a water bath for 10 minutes. 2 $\mu$g of total cell extracts were separated on homogeneous 12.5% PhastGel (Phastsystem, Amersham Pharmacia Biotech, NJ, USA) with SDS buffer strips according to manufacturers instructions with subsequent transfer to PVDF membrane for 40 minutes. The membrane was blocked by 0.5% blocking solution (skim milk-based), incubated with anti-Bcl-2 antibody, anti-Bad antibody, anti-Bcl-x antibody, anti-Bax antibody (Transduction Laboratories, Calif, USA), anti-CD47 antibody (Lab Vision, Calif, USA), or anti-$\beta$-actin antibody (Chemicon Interna-tional, Calif) at 25°C followed by incubation with peroxidase-coupled secondary antibody (Kirkegaard Perry Laboratories, Md, USA) and detected by the ECL-enhanced chemiluminescent system (Boehringer Mannheim, Germany).

## 3. RESULTS

### 3.1. *Kinetics of gene expression of HBMEC showing the progression of gene profile changes during C. neoformans infection*

Microbial invasion is a complex and dynamic process. We performed a time-course study to examine the gene

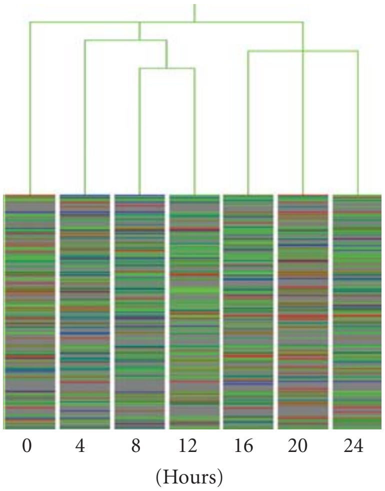


Figure 1: Gene expression profiles during *C. neoformans* infection; microarray analysis of mRNA levels was assessed at 0, 4, 8, 12, 16, 20, and 24 hours after *C. neoformans* incubation. Time points are represented by columns, and genes in rows. *Red, green*, and *blue* represent the higher, equal, and lower mRNA level relative to that of zero time point. A hierarchical clustering analysis of genes with expression levels that changed during *C. neoformans* and HBMEC interactions was performed. The distance among infectomic profiles of HBMEC is shown. Each lane represents a result form a gene chip. Genes with high intensity reading are shown in bright colors; genes with low intensity reading are in dim colors (grey) so they can be distinguished during analysis.

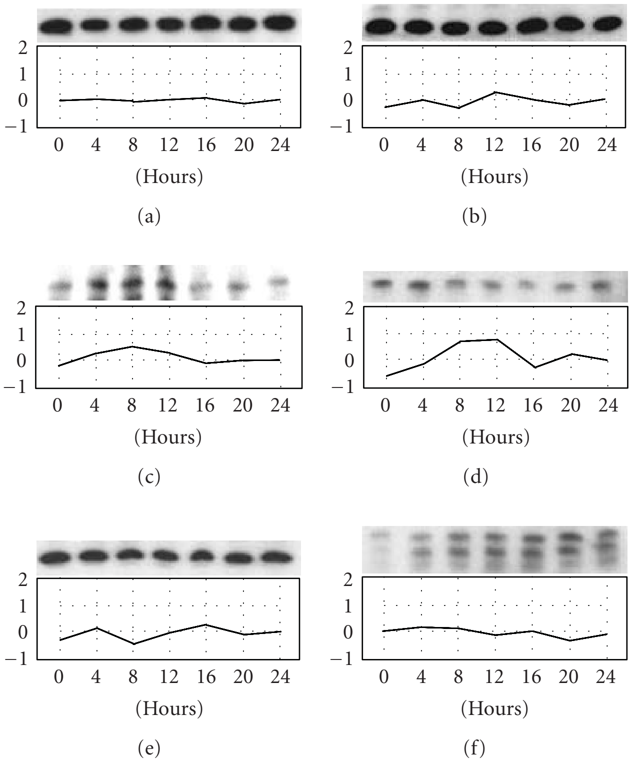


Figure 2: Comparison of results obtained with microarray and protein blot analysis for changes in protein levels during *C. neoformans* infection. Protein levels at various time points (*top panel*) relative to the mRNA levels in microarray analysis (*bottom panel*) are shown for the following genes: (a) $\beta$-actin, (b) Bcl-x, (c) CD47, (d) Bax, (e) Bad, and (f) Bcl-2. The relative scales of the mRNA expression are indicated on the $y$-axis.

expression profiles in HBMEC during *C. neoformans* infection. We selected *C. neoformans* strain B3501 for this study because the genomic sequence of this strain has been completed. In addition, we have used this strain for in vitro adhesion and transcytosis studies on HBMEC [11, 13]. It is an encapsulated strain with moderate adhesion activity to HBMEC, which was used as the in vitro model of BBB. A spectrum of 24 hours would allow us to evaluate its effect on HBMEC. Poly(A)$^+$ RNAs derived from the incubated HBMEC at 0, 4, 8, 12, 16, 20, or 24 hours after B3501 incubation were subjected to the RNA extraction and the preparation of the biotinylated cRNA probe. The prepared probes at different time points were hybridized with biochip, individually. We used an Affymetric HU95A microarray chip harboring 12,559 human clones, facilitating efficient detection of changes in gene expression in the host cells. The expression levels were measured and analyzed by the GeneChip program. In a hierarchical clustering, the closet pair of expression values is grouped and the data points are organized in a phylogenetic tree in which the branch lengths represent the degree of similarity between the values. The mRNA profiles of HBMEC, assessed at 0, 4, 8, 12, 16, 20, and 24 hours after *C. neoformans* incubation, are shown in Figure 1.

These 12,559 gene patterns were classified into two major clusters (Clusters 0–12 hours and 16–24 hours) by using the GeneSpring software. Rapid changes in gene profiles at an early time point are followed by a gradual alteration until the 12-hour time point. Of interest, the gene patterns of the last three time points were similar, yet more distal related to the early gradual changes of gene profiles (0 to 12 hours). At these time points (16, 20, 24 hours), the gene profile seems to reach a plateau. The HBMEC is a homogeneous cell culture. The changes in the gene expression profiles are not due to the heterogeneity of cell cultures. The perturbation of gene expression profiles is most likely due to the presence of *C. neoformans*. Overall, the hierarchical clustering analysis of the mRNA levels reveals a sequential change in HBMEC infected with *C. neoformans*. The overall profile was altered more prominently at the initial stage of the pathogen-host interaction and subsequently persistent expression.

### 3.2. Comparison of gene expression profiles from the microarray and protein levels from Western blot analyses

To determine the validity of results obtained by the microarrayanalysis, some genes were subjected to Western analysis. These genes included $\beta$-actin, Bcl-x, CD47, Bax, Bad, and Bcl-2 (see Figure 2). They were chosen for blot analysis because their protein levels were variable and the affinity of commercially available antibodies. For example, $\beta$-actin mRNA and protein levels are maintained in a constant level

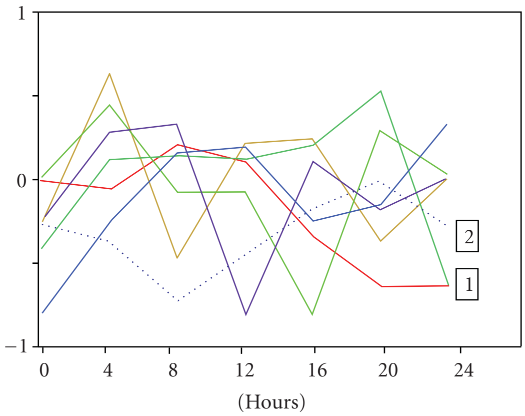


Figure 3: PCA of HBMEC profiles after *C. neoformans* infection; microarray data were analyzed by the GeneSpring software to classify the gene expression patterns. Seven major gene profiles were obtained as shown in different color. The major groups (PCA groups no.1 and no.2) are indicated in the graph, which contain more than 50% of total analyzed genes.

at all time points, whereas CD47 mRNA and protein level increase to the 8 hour and then decline back to the normal levels. In general, the protein levels of these genes were expressed in varied degrees, and the protein expression profiles were comparable between microarray and protein analyses (see Figure 2). The results validated the competency of the microarray analysis for detection of changes in gene expression in HBMEC during *C. neoformans* infection.

### 3.3. Principal component analysis (PCA) of HBMEC gene profiles during C. neoformans infection

In order to obtain a global grouping of gene profiles, we examined the PCA grouping of mRNA from HBMEC. PCA was a useful linear approach for obtaining a simplified visualization of entire datasets, without losing experimental information (variance). PCA allowed the dimension of complex data to be reduced and the most relevant features of a given dataset (transcriptome) to be highlighted. It was useful for our kinetic studies because it was possible to describe trends that were, otherwise, irretrievably by a direct examination of the entire dataset. The gene fluctuation could be classified into 7 major groups. The major group (PCA group 1) contained 4155 genes (33.08% of total analyzed genes) and the PCA group 2 contained 3129 genes (24.92% of total analyzed genes) (see Figure 3). Thus, these two groups represented more than 50% of HBMEC mRNAs. The gene profile of PCA group 1 showed a very slightly increase at 8 and 12 hours, and then a rapid decline. On the other hand, the PCA group 2 showed a gradual decline till 8 hours and then bounced back to near the original levels.These two profiles were nearly a mirror image.The results suggested that HBMEC altered its mRNA levels or adjusted its physiological status in response to the pathogen invasion.

### 3.4. Gene expression profiles at different time points in HBMEC analyzed by the self-organizing maps (SOM)

The interpretation of system complexity is the most challenging task of biology in this century. The analysis of complexity in biological systems might start from a simplified representation of static gene networks and then move to an increasingly well-defined and integrated description of biological phenomena, bearing in mind that only dynamic networks will explain reality adequately. An example of a nonhierarchical clustering method is SOM. As SOM solves difficult high-dimensional and nonlinear problems. Using this algorithm, gene expression data are transformed into vectors or coordinates in an *n*-dimensional space, where *n* equals the number of variables or time points. Gene profiles of HBMEC postinfection at 7 time points (0, 4, 8, 12, 16, 20, 24 hours) were analyzed by SOM (7 × 6) (see Figure 4). Upper bound (4.0), normal (1.0), and lower bound (0.0) were shown. The dynamic of gene expression during pathogen-host interaction can be clearly observed (see Figure 4). The use of SOM for grouping of different profiles greatly facilitates further analysis.

### 3.5. Persistently upregulated and down-regulated genes during C. neoformans infection

Despites significant fluctuation of gene expression profiles in most HBMEC genes in response to *C. neoformans* infection, there were some genes that were upregulated or downregulated persistently during 24 hours period. One example of the upregulated genes was the cytochrome P450 gene which was continuously upregulated through the course of infection. A similar gene profile could be found and grouped, for example, genes that had a correlation at least 0.95 to the expression profile of cytochrome P450 were defined as upregulated genes (see Figure 5(a)). In the same manner, downregulated genes were grouped, in which genes that had a correlation at least 0.95 to the expression profile of troponin I type 3 gene TNNI3 (see Figure 5(b)).

### 3.6. Induced fluctuation of interferons and MHC group genes during C. neoformans infection

We have performed the hierarchical and gene tree analyses to test whether the interferon genes and MHC-related genes were fluctuation during *C. neoformans* infection. The group of interferon-related genes (66 genes) was clustered based on their expression level (see Figure 6). Most genes were fluctuated during the course of infection, suggesting interferon played a role in the innate immune response to *C. neoformans*. Many infectious microbes induced expression of chemokines and adhesion molecules in human endothelial cells. Previous studies of the expression of IL-8, INF-$\gamma$-inducible protein-10 (IP-10), MCP-1, and the leukocyte ligand ICAM-1 in primary HUVEC revealed that *C. neoformans* had the ability to interfere with inflammatory signaling in human endothelial cells, and suggested that *C. neoformans* might induce leukocyte activation and trafficking in

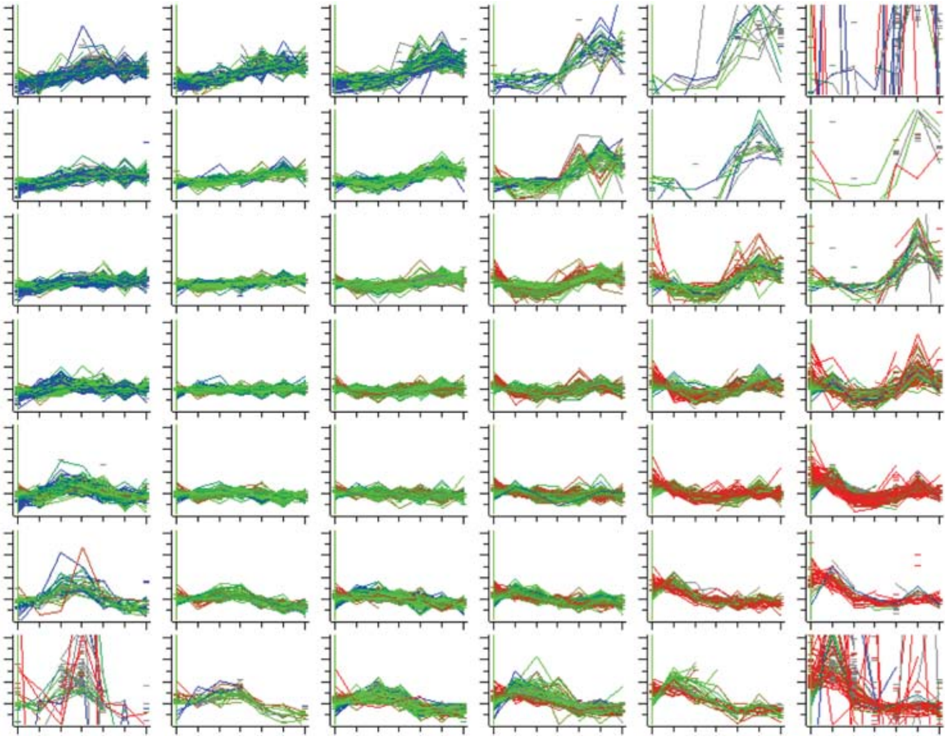
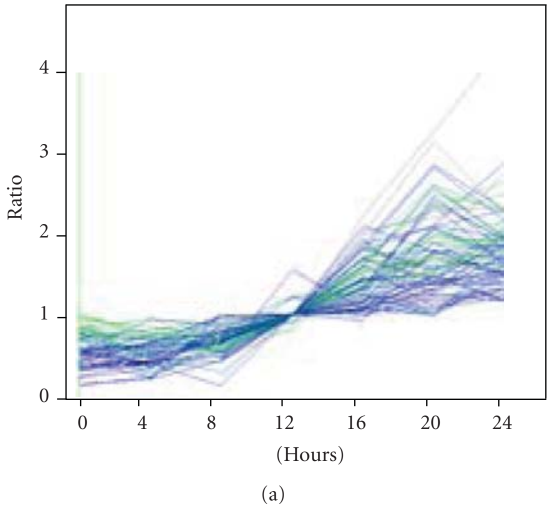


Figure 4: Infectomic profiles of *C. neoformans*-infected HBEMC using SOM analysis. Gene profiles of HBMEC postinfection at 7 time points (0, 4, 8, 12, 16, 20, 24 hours) ($x$-axis) were analyzed by SOM (7 x 6). Upper bound (4.0), normal (1.0), and lower bound (0.0) were shown ($y$-axis).

the infected host (HUVEC) [5, 6]. Our results showed that IL-1 increased slightly, while MCP-1 did not change significantly. No changes in IP-10 and ICAM-1 expressions were observed in this chip. The TNF-$\alpha$ profile was also increased, though quite late, during the infection. Similarly, the expression of a group of MHC genes including 29 of those in MHC II class was fluctuated (see Figure 7). Although some genes were listed but not expressed, our results suggested that endothelial cells contributed to the host immune response to *C. neoformans* infection. Taken together, HBMEC is not a professional immune cell; however, the fluctuation in the expression of interferons and MHC genes suggests that the innate immune systems are activated during the pathogen invasion.

## 4. DISCUSSION

The current work demonstrates the use of cDNA microarray-based infectomic approach to characterize transcriptomes in HBMEC infected with the meningitic pathogen *C. neoformans*. Like most of meningitic pathogens, the penetration across the BBB that is constituted by BMEC is required for the pathogenesis of the CNS infection caused by *C. neoformans* [6, 14]. Our study represents the first investigation of the holistic transcriptional response of the in vitro BBB cell system in response to *C. neoformans* infection. A total of 12,559 human genes have been analyzed in this study, and distinct alterations in HBMEC gene expression have been observed at 4, 8, 12, 16, 20, and 24 hours of infection. Significant changes in the transcriptional infectomes were observed in HBMEC infected with *C. neoformans*. This approach was used to define the infectomic profiles of HBMEC infected with *C. neoformans* for a series of time points, thus, evaluating the dynamic changes in gene expression profiles and inflammatory factors in the host response to this fungal pathogen. The data were analyzed with several clustering analyses. In a hierarchical clustering analysis, the gene expression patterns at different time points displayed a progressive change of gene profiles. The first three incubated time points were clustered, and showed more drastic changes at initial engagement between *C. neoformans* and HBMEC (4 hours). The gene partners of the latter time points (16, 20, 24 hours) were clustered together, indicating the alternations reached to a plateau. Microarray analysis has been used for monitoring host gene expression profiles from other pathogen invasion studies, such as HIV-1. Many of the results were to compare the data pre- and post- infections. Our studies revealed that pathogen invasions are multifaceted and dynamics. A time point study is necessary to monitor the alterations. In a nonhierarchical analysis SOM, the two-dimensional space map displayed different expression profiles form 12,559 genes into 42 groups. One interesting finding defined by PCA is that one group of genes with graduate reduction in expression (∼33% of total genes) accompanied with another group of genes showing gradually increased expression profile (∼25% of total genes). The alternative changes in the two major groups (1 and 2) are shown as a mirror image (see Figure 3). The biological mechanism and significance of the expression profile changes in



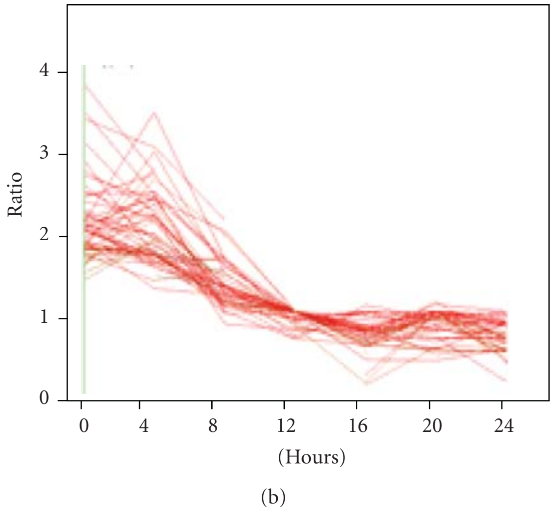


Figure 5: The upregulated and downregulated gene profiles of HBMEC during *C. neoformans* infection. (a) Upregulated genes (left chart): cytochrome P450 gene was continuously upregulated during the course of infection. Genes that have correlation at least 0.95 to the expression profile of cytochrome P450 are grouped and defined as upregulated genes. (b) Downregulated genes (right chart): TNNI3 gene was continuously downregulated during the course of infection. Genes that have correlation at least 0.95 to the expression profile of TNNI3 are grouped and defined as downregulated genes.

the pathogenesis of *C. neoformans* infection remain to be defined.

Most interestingly, the group of interferon-related genes (66 genes) was clustered based on their expression level. The expression of most genes fluctuated during the course of infection, suggesting that interferons play a role in the innate immune response to the pathogen. The expression of 29 MHC II class also fluctuated. Though some genes were listed but not expressed, the results suggested that endothelial cells contributed to the host immune response to *C. neoformans* infection. Many infectious microorganisms induced expression of chemokines and adhesion molecules in human endothelial cells. Studies of the gene expression in primary HUVEC revealed that *C. neoformans* had the ability to interfere with inflammatory signaling in human endothelial cells, and suggested that *C. neoformans* may later induce leukocyte activation and trafficking in the infected host (HUVEC) [15, 16]. Our results also showed that expression of interferon-related genes and MHC group genes are changed during *C. neoformans* infection. Collectively, the fluctuation in the expression of innate-related genes suggests that *C. neoformans*, like other microbial pathogens, is able to induce expression of chemokines and other innate genes.

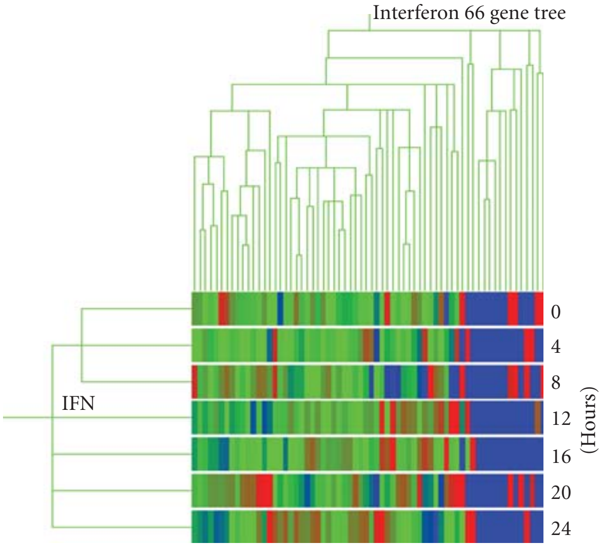


Figure 6: Hierarchical cluster of interferon 66 gene tree. Different expression of 66 interferon-related genes were analyzed by Gene Tree program. The expression of more than two thirds (red and blue) was changed, and about one third of the genes were not expressed (green).

The BBB has been considered an immunologically inactive organ as there are few antigen-presenting cells present in the CNS and the presence of the tight junction was thought to prevent the entry of immune cells from the peripheral circulations into the CNS [8]. However, increasing number of studies indicate that the endothelium of the BBB constitutes a dynamic and immunoactive interface between the blood and the CNS that can be modulated by endogenous factors such as bradykinin and cytokines, as well as exogenous factors including meningitic pathogens and their products [17]. Alterations in the BBB function are critical for the development of CNS infection including cryptococcal meningoencephalitis. It has been learnt about the immune response to cryptococcal infection from both in vitro and in vivo studies with the focus on immune cells [6, 18]. The immune response to *C. neoformans* infection was observed as one of the most interesting changes.

## 5. CONCLUSION

In summary, the current infectomic studies with a genome survey of *C. neoformans*-induced host response in HBMEC provide global information for the pathogenesis of the CNS infection caused by this fungus. Similar conclusions

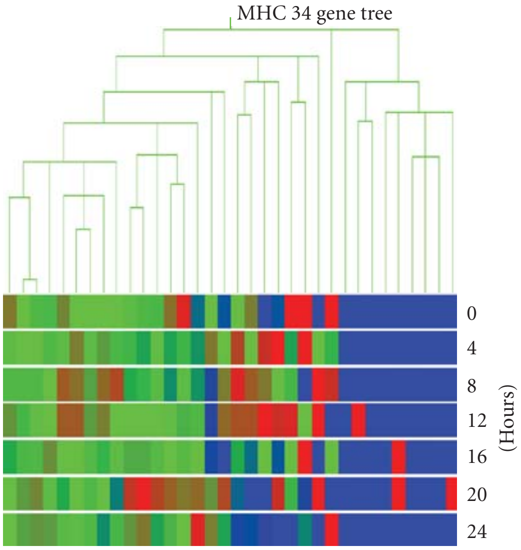


Figure 7: Gene tree of MHC genes fluctuated during *C. neoformans* infection: expression profiles changes in HBMEC are observed. A subgroup of the profile with MHC class genes was selected and subjected to Gene Tree analysis using GeneSpring$^{TM}$ software. The alternations of gene expression in this group suggested that innate immune contributes to *C. neoformans* infection.

were reached when analyzing the microarray data with two complementary approaches, PCA and SOM. However, PCA works well in problem spaces that are linearly separable. SOM solves difficult high-dimensional and nonlinear problems. Most importantly, our study demonstrates for the first time that the BBB can actively participate in immune response to cryptococcal infection by regulating the expression of interferons, MHC and cytokines. It suggests that the BBB is not only an impermeable cellular barrier but also constitutes an immunoactive interface between the circulatory system and the CNS and that can be modulated by meningitic pathogens such as *C. neoformans*. Further insight into the role of the BBB in the pathogenesis of microbial meningitis and immunoregulation of the host defense against meningitic pathogens will gain a global view of the CNS infection and offer exciting prospects for advances in the prevention and therapy of this disease.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. K. J. Kwon-Chung and Dr. Y. Chang for providing *C. neoformans* strains and intellectual inputs, and Dr. Kwang S. Kim for providing HBMEC. They also thank their colleagues for suggestions, help, and discussion. This work was supported in part by grants-in-aid from the NIH Grant no. RO1 NS047599 (A.Y.J.) and Grant no. RO1 AI40635 (S.H.H.).

## REFERENCES

[1] S.-H. Huang, T. Triche, and A. Jong, “Infectomics: genomics and proteomics of microbial infections,” *Functional & Integrative Genomics*, vol. 1, no. 6, pp. 331–344, 2002.

[2] T. G. Mitchell and J. R. Perfect, “Cryptococcosis in the era of AIDS—100 years after the discovery of *Cryptococcus neoformans*,” *Clinical Microbiology Reviews*, vol. 8, no. 4, pp. 515–548, 1995.

[3] M. Gottfredsson and J. R. Perfect, “Fungal meningitis,” *Seminars in Neurology*, vol. 20, no. 3, pp. 307–322, 2000.

[4] J. R. Perfect, B. Wong, Y. C. Chang, K. J. Kwon-Chung, and P. R. Williamson, “*Cryptococcus neoformans*: virulence and host defences,” *Journal of Medical and Veterinary Mycology*, vol. 3, supplement 1, pp. 79–86, 1998.

[5] M. Feldmesser, Y. Kress, P. Novikoff, and A. Casadevall, “*Cryptococcus neoformans* is a facultative intracellular pathogen in murine pulmonary infection,” *Infection and Immunity*, vol. 68, no. 7, pp. 4225–4237, 2000.

[6] T. Bicanic and T. S. Harrison, “Cryptococcal meningitis,” *British Medical Bulletin*, vol. 72, no. 1, pp. 99–118, 2005.

[7] G. B. Huffnagle and L. K. McNeil, “Dissemination of C. neoformans to the central nervous system: role of chemokines, Th1 immunity and leukocyte recruitment,” *Journal of NeuroVirology*, vol. 5, no. 1, pp. 76–81, 1999.

[8] L. L. Rubin and J. M. Staddon, “The cell biology of the blood-brain barrier,” *Annual Review of Neuroscience*, vol. 22, pp. 11–28, 1999.

[9] R. D. Broadwell, B. J. Baker-Cairns, P. M. Friden, C. Oliver, and J. C. Villegas, “Transcytosis of protein through the mammalian cerebral epithelium and endothelium. III. Receptor-mediated transcytosis through the blood-brain barrier of blood-borne transferrin and antibody against the transferrin receptor,” *Experimental Neurology*, vol. 142, no. 1, pp. 47–65, 1996.

[10] S.-H. Huang and A. Jong, “Cellular mechanisms of microbial proteins contributing to invasion of the blood-brain barrier,” *Cellular Microbiology*, vol. 3, no. 5, pp. 277–287, 2001.

[11] S. H. M. Chen, M. F. Stins, S.-H. Huang, et al., “*Crytptococcus neoformans* induces alterations in cytoskeleton of human rain microvascular endothelial cells,” *Journal of Medical Microbiology*, vol. 52, pp. 961–970, 2003.

[12] Y. C. Chang, A. Jong, S.-H. Huang, P. Zerfas, and K.J. Kwon-Chung, “*CPS1*, a homolog of the *Streptococcus pneumoniae* type 3 polysaccharide synthase gene, is important for the pathobiology of *Cryptococcus neoformans*,” *Infection and Immunity*, vol. 74, no. 7, pp. 3930–3938, 2006.

[13] A. Jong, M. F. Stins, S.-H. Huang, S. H. M. Chen, and K. S. Kim, “Traversal of *Candida albicans*across human blood-brain barrier in vitro,” *Infection and Immunity*, vol. 69, no. 7, pp. 4536–4544, 2001.

[14] K. E. Black and L. R. Baden, “Fungal infections of the CNS: treatment strategies for the immunocompromised patient,” *CNS Drugs*, vol. 21, no. 4, pp. 293–318, 2007.

[15] T. R. Traynor and G. B. Huffnagle, “Role of chemokines in fungal infections,” *Medical Mycology*, vol. 39, no. 1, pp. 41–50, 2001.

[16] N. Mozaffarian, A. Casadevall, and J. W. Berman, “Inhibition of human endothelial cell chemokine production by the opportunistic fungal pathogen *Cryptococcus neoformans*,” *Journal of Immunology*, vol. 165, no. 3, pp. 1541–1547, 2000.

[17] A. A. Siddiqui, A. E. Brouwer, V. Wuthiekanun, et al., “IFN-$\gamma$ at the site of infection determines rate of clearance of infection in cryptococcal meningitis,” *Journal of Immunology*, vol. 174, no. 3, pp. 1746–1750, 2005.

[18] J. R. Graybill, “Cryptococcal meningitis,” *Brazilian Journal of Infectious Diseases*, vol. 1, pp. 60–67, 1997.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 547428, 10 pages
doi:10.1155/2008/547428

*Methodology Report*

# A Suprachoroidal Electrical Retinal Stimulator Design for Long-Term Animal Experiments and In Vivo Assessment of Its Feasibility and Biocompatibility in Rabbits

**J. A. Zhou,[1, 2, 3] S. J. Woo,[1, 2, 4, 5] S. I. Park,[1] E. T. Kim,[1, 2, 3] J. M. Seo,[1, 2, 6] H. Chung,[1, 2, 4] and S. J. Kim[1, 2, 3]**

[1] *Nano Bioelectronics & Systems Research Center, Seoul National University, Shillim-dong, Gwanak-gu, Seoul 151-742, South Korea*
[2] *Nano Artificial Vision Research Center, Seoul National University Hospital, Yeongeon-dong, Jongno-gu, Seoul 110-744, South Korea*
[3] *School of Electrical Engineering and Computer Science, Seoul National University, Shillim-dong, Gwanak-gu, Seoul 151-742, South Korea*
[4] *Department of Ophthalmology, Seoul National University College of Medicine, Yeongeon-dong, Jongno-gu, Seoul 110-799, South Korea*
[5] *Seoul National University Bundang Hospital, Gumi-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 463-707, South Korea*
[6] *Department of Ophthalmology, Dongguk University College of Medicine, Pil-dong, Jung-gu, Seoul 100-715, South Korea*

Correspondence should be addressed to S. J. Kim, kimsj@snu.ac.kr

Received 21 August 2007; Revised 29 November 2007; Accepted 24 December 2007

Recommended by Daniel Howard

This article reports on a retinal stimulation system for long-term use in animal electrical stimulation experiments. The presented system consisted of an implantable stimulator which provided continuous electrical stimulation, and an external component which provided preset stimulation patterns and power to the implanted stimulator via a paired radio frequency (RF) coil. A rechargeable internal battery and a parameter memory component were introduced to the implanted retinal stimulator. As a result, the external component was not necessary during the stimulation mode. The inductive coil pair was used to pass the parameter data and to recharge the battery. A switch circuit was used to separate the stimulation mode from the battery recharging mode. The implantable stimulator was implemented with IC chips and the electronics, except for the stimulation electrodes, were hermetically packaged in a biocompatible metal case. A polyimide-based gold electrode array was used. Surgical implantation into rabbits was performed to verify the functionality and safety of this newly designed system. The electrodes were implanted in the suprachoroidal space. Evoked cortical potentials were recorded during electrical stimulation of the retina. Long-term follow-up using OCT showed no chorioretinal abnormality after implantation of the electrodes.

## 1. INTRODUCTION

Retinal prostheses are under investigation by several groups [1, 2], and some preclinical and clinical trials have been reported [3–6]. Preclinical experiments are intended to estimate the stimulation parameters and to evaluate the efficacy and the safety of the devices and the clinical trials are aimed at demonstrating the feasibility of the prosthesis. Although the electrical stimulation of the retina in preliminary clinical studies showed encouraging results such as the patient's perception of a small spot of light or basic shapes according to the stimulation pattern [4], there is much to be investigated and revised regarding retinal prostheses.

Before the implantation of a retinal prosthesis into the eye of a patient, the stimulation conditions, the long-term stability, and the durability of the retinal prostheses should be verified in vivo. Animal experiments are essential in developing retinal prosthesis because the characteristics and safety of prosthesis cannot be verified in human eyes for ethical reasons. Due to anatomical differences, devices for animal use cannot be used in humans, but design features can be tested in extended animal experiments, verifying the design features for a human device.

Previous reports showed long-term biocompatibility of various electrodes without electrical stimulation of the retina [7–10]. Thresholds of electrical stimulation were presented

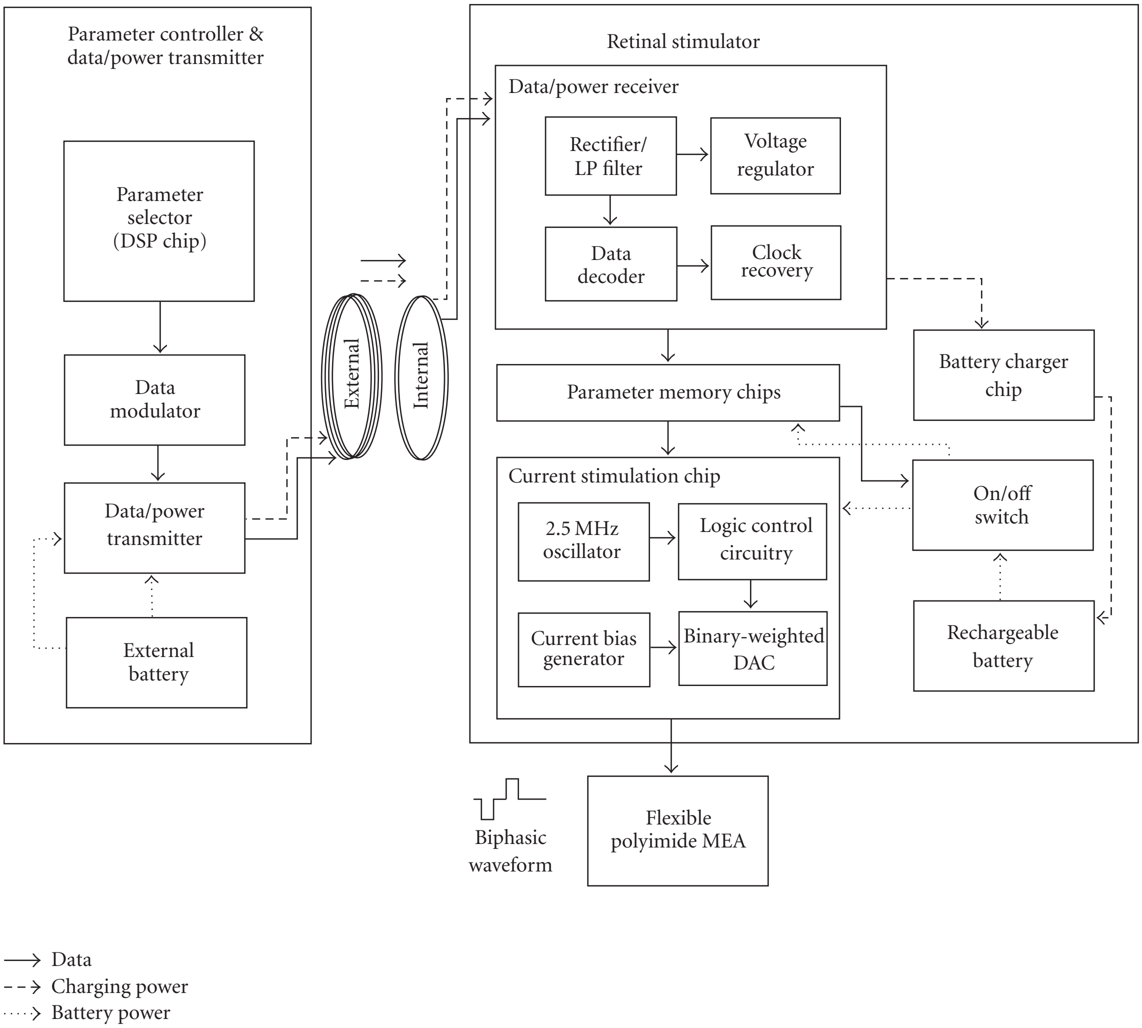


Figure 1: System block diagram.

from acute implantation experiments [11–13] and also from chronic human studies [14, 15]. In previous studies, electrical stimulation systems needed a wired or a wireless external component to provide stimulation parameters and power to the internal electrodes. When those systems were used for long-term electrical stimulation in animals [16, 17], there were many problems that needed to be solved. Systems with a percutaneous connection to the external portion restricted the animals' movement and posed an infection risk. In systems having transcutaneous connections to an external controller worn by the animals, the external part usually separated from the animals or was damaged.

In this article, an implantable retinal prosthesis system is proposed for a chronic electrical stimulation test in an animal model. For this purpose, a small rechargeable battery and a parameter memory were introduced into the implanted stimulator so the external power supply and control part could be removed during a chronic stimulation experiment. The animal is then free of any external components. The external unit is then needed temporarily for two purposes only: passing the parameter and charging the battery. To our knowledge, this is the only retina stimulation system designed for use in animals with such a feature.

Animal experiments were done to show the feasibility of the implantation of this newly suggested stimulation system. To check whether the implanted electrode could induce appropriate cortical response upon electrical stimulation, we measured the electrically evoked cortical potentials (EECPs) from rabbits in which electrodes were implanted. The long-term biocompatibility of electrodes was also evaluated in vivo with OCT.

## 2. METHODS

### 2.1. *Retinal prosthesis system design*

The implantable retinal prosthesis system for a chronic animal experiment consisted of an internal unit for retinal stimulation and an external unit for stimulation control and battery charging (see Figure 1). A paired RF coil links these two units for the transmission of data and power.

MSB

| | | | | | | |
|---|---|---|---|---|---|---|
| *S* (1-bit) | *C*2–0 (3-bit) | *D*3–0 (4-bit) | *R*3–0 (4-bit) | *A*7–0 (8-bit) | *P* (1-bit) | EOF (1-bit) |

*S*: Stimulation on/off; $S = 1$ on; $S = 0$ off

*C*2–0: Electrode selection; up to 7 channels

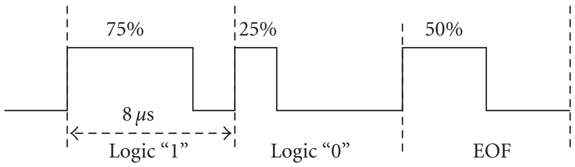
*D*3–0: Duration setting; up to 3 ms with 200 $\mu$s resolution

*R*3–0: Period setting; up to 250 ms with 16 ms resolution

*A*7–0: Amplitude setting; up to 2 mA with 8 $\mu$A resolution

*P*: Odd parity check bit;

EOF: End of frame



PWM modulation

Figure 2: Data formats and PWM modulation.

The external unit had a stimulation waveform parameter selector used to control the channel selection, amplitude, duration, and rate of stimulation. This parameter selector generated a parameter data frame and was implemented using a commercially available digital signal processing chip (TMS320VC5509, Texas Instruments, Dallas, Tex, USA). The control codes were implemented inhouse using the C programming language, and the parameter data frame consisted of 22 bits as shown in Figure 2. The same stimulation waveform was simultaneously delivered to all selected channels. To transmit this parameter data into the internal stimulator, the pulse width modulation (PWM) encoding method was used (see Figure 2). Logic “1” and “0” were encoded to have a duty cycle of 75% and 25%, respectively, and the “end-of-frame (EOF)” bit had a 50% duty cycle. Such an encoding method enables easier synchronization and decoding because each bit had a uniform rising edge at its beginning [18]. The transmission data rate was 125 kbps. For transmission of PWM encoded data, a class-E tuned power amplifier (data/power transmitter) was used with amplitude shifted keying (ASK) modulation. The carrier frequency was 2.5 MHz.

The transmitted data were received by the internal coil and then the envelope was extracted through a half-wave rectifier and a lowpass filter. Using this envelope signal, a data decoder in the data/power receiver chip recovered the parameter data and saved it to the parameter memory chip. Using the same envelope signal, a voltage regulator generated power to be consumed by the data/power receiver chip (see Figure 1).

The internal unit, that is, the retinal stimulator, was implemented with a rechargeable battery and integrated circuit (IC) chips including the data/power receiver chip, current stimulation chip, parameter memory chips, and battery charging chip. The data/power receiver chip had data decoding and voltage regulation function blocks. Both the data/power receiver chip and the current stimulation chip are custom IC’s designed by our laboratory (0.8 $\mu$m complementary metal-oxide semiconductor (CMOS) technology, Austria Microsystems, Unterpremstaetten, Austria). The parameter memory and battery charging chips were off-the-shelf commercial products (see Figure 1). Except for the data/power receiver chip, the other chips in the stimulator were powered by a rechargeable battery. Therefore, once the parameters were passed to the parameter memory, the external unit can be removed from the animal during the electrical stimulation test. The retinal stimulator had two modes of function: a stimulation mode and a battery recharging mode.

In stimulation mode, the saved parameter data in the parameter memory component were provided to the current stimulation chip. The parameter memory component was composed of three 8-bit shift registers (SN54AHC595, Texas Instruments, Dallas, Tex, USA) for 22-bit data storing. The parameter data did not change unless a new parameter was transmitted from the external component. The current stimulation chip consisted of seven current sources and a timing logic circuitry. The current generator circuitry had current bias circuitry (8 $\mu$A) and an 8-bit binary current-weighted DAC (digital-to-analog converter). The timing logic circuitry had a 2.5 MHz oscillator and switch control logic circuitry for controlling the current stimulation waveform.

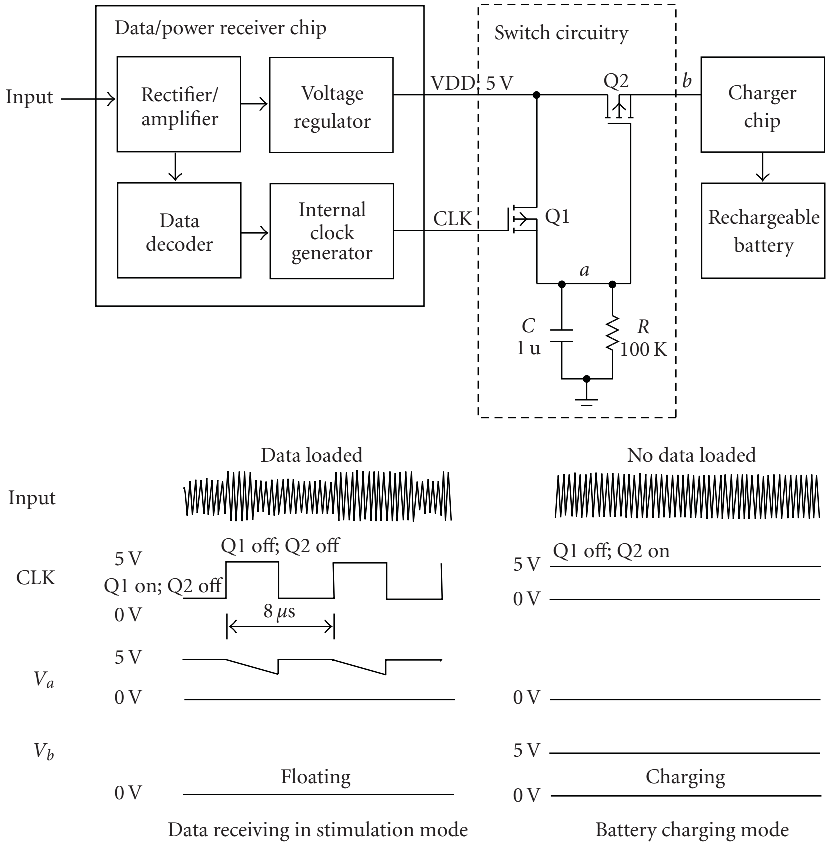


Figure 3: Operation of switch circuitry in stimulation and battery charging modes. When no data were loaded on 2.5 MHz carrier, CLK was logic high, therefore switch Q1 would be turned off, and the voltage of "a" node would be logic low, so switch Q2 would be turned on (battery recharging mode). If any data were loaded on 2.5 MHz carrier, the CLK would recover and the voltage of node "a" never went below the threshold of switch Q2, so Q2 was turned off, this would be disabling the charger chip period (stimulation mode).

In the battery recharging mode, 2.5 MHz sinusoidal waves were transmitted with no data. A rechargeable coin-type lithium ion battery (PD2320, Korea Power cell Inc., Daejeon, Korea) was used as the power source for the internal implant. A charging chip (LTC4054L, Linear Technology Corporation, Milpitas, Calif, USA) was used to control the battery recharging.

In this work, only one inductive coupling was used for data transmission and battery charging. Simultaneous transmission of the stimulation parameter and charging power is difficult because the battery charger circuit affects the precisely designed load value of the data/power receiving circuit and can induce the failure in data reception. To separate the stimulation mode and battery charging mode, a switch circuit was positioned between the voltage regulator in the data/power receiver chip and the battery charge chip to control the recharging of the battery. The introduced switch circuit consisted of two p-MOS transistors, one capacitor and one resistor (see Figure 3). The resistor and capacitor comprised a parallel connection with an RC time constant of 100 millisecons, which was very long compared to the 8-microsecond period of the clock of the data/power receiver chip. Therefore, the voltage of the "a" node was higher than the threshold of the Q2 switch when a data signal (PWM) was applied causing the Q2 to turn off. The data decoding could therefore be successfully carried out with no load effect. However, when only a sinusoidal wave without data was applied to the retinal stimulator through the inductive link, the level of CLK was logic high (see Figure 3). In this case, Q1 would be turned off and the voltage of node "a" would be logic low, so Q2 would be turned on. Therefore, the battery could be recharged.

To protect the ICs from body fluids and mechanical forces, the electronics of the stimulator were hermetically housed in a metal package which consists of biocompatible titanium housing, platinum feedthroughs, and a ceramic plate. The feedthroughs connected the electrode array and receiver coil to the retinal stimulator. A ceramic sintering process was used to fix the feedthroughs in the ceramic plate that provided electrical isolation. Brazing and laser welding techniques were employed to achieve hermetic sealing of the titanium housing [18].

A polyimide-based seven-channel, strip-shaped (750 × 300 μm) gold electrode array was used as stimulating electrode. The stimulation sites were constructed in a 4 mm × 4 mm area with a seven segmented configuration [17]. A large circle electrode in 1500 μm diameter, also polyimide-based, was used as the reference electrode (see Figure 4).

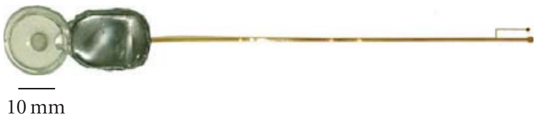


Figure 4: A photograph of retinal electrical stimulator which consists of receiver coil, hermetically sealed metal package, and polyimide-based active and reference electrodes.

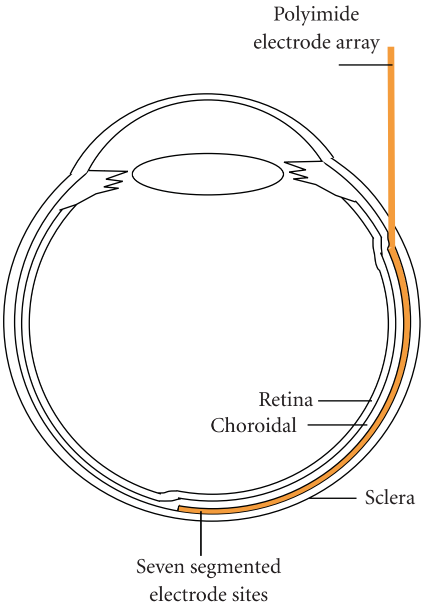


Figure 5: A diagram showing the positioning of an electrode array in the rabbit eye ball. The electrode array is colored in yellow. (Figure 9(a) shows a fundus photography of rabbit retina with a suprachoroidal electrode.)

The thickness of polyimide/gold electrode array was 58 $\mu$m. One side of this array was lengthened and connected to the stimulator via feedthroughs.

### 2.2. *Surgical procedures for implantation of the prosthesis system into rabbits*

New Zealand White rabbits weighing 2.0 ~ 2.5 kg were used to evaluate the proposed system. All procedures of animal experimentation were approved by the Institutional Review Board of Seoul National University Hospital Clinical Research Institute and followed the Association for Research in Vision and Ophthalmology (ARVO) Statement on Use of Animals in Ophthalmic and Vision Research. Implantation of the entire system was performed under general anesthesia achieved by repetitive intramuscular injection of 25 mg ketamine and 6 mg xylazine per kilogram of body weight.

The skin was prepared and a longitudinal incision was made at the right auscultation triangle of the back. Meticulous dissection was done between the subcutaneous and muscular layers. A subcutaneous tunnel was made by an elevator from the medial angle of the scapula to the inferior conjunctival fornix. The forniceal opening was created with a blade and the tip of the elevator was extruded from the forniceal opening thereby completing the subcutaneous tunnel. The internal stimulator was inserted into the subcutaneous tunnel from the opening of the back. The connection part and the polyimide electrode array were protected by a soft polyethylene tube and passed through the tunnel. After being introduced through the conjuctival forniceal opening, the connection wire was turned around the eyeball under the extraocular muscles and temporarily anchored onto the sclera with two 6-0 Prolene sutures, which is a process similar to Humayun's method [4].

A 5 mm sized Scleral tunnel incision parallel to the limbus was made with a blade and scissors about 5 mm away from the limbus in the upper lateral quadrant of the eyeball. To prevent vitreous prolapse and to lower the intraocular pressure, 0.1–0.2 mL of aqueous humor was drained from the anterior chamber before the scleral opening was created. The polyimide electrode array was inserted through the scleral opening and slided into the suprachoroidal space to reach the visual streak. Figure 5 shows the position of the electrode array in the rabbit eye ball. The scleral opening was closed with 8-0 Vicryl sutures and the connection wire was permanently anchored onto the sclera with sutures. The reference electrode was left outside the eyeball to contact the sclera without fixation. The conjunctival incisions were repaired with 8-0 Vicryl sutures and the skin incision was repaired with 6-0 Catgut sutures. After the operation, the entire system was inside the body and was not exposed.

### 2.3. *Recording of visually/electrically evoked cortical potentials*

Stainless needle electrodes were used as recording electrodes. An active recording electrode was placed into the primary visual cortex 6 mm lateral and 6 mm anterior to lambda, which is the same location as in Okuno's method [19]. A reference recording electrode was placed into the cortex 20 mm anterior to lambda. The animal was grounded by an electrode on the ipsilateral ear. The visually evoked cortical potentials (VECPs) and electrically evoked cortical potentials (EECPs) elicited by stimulating one eye were recorded from the active recording electrode placed on the contralateral side.

An integrated hardware/software platform (TDT System 3, Tucker-Davis Technologies, Alachua, Fla, USA) was used for amplifying, acquisition, and storage of VECP and EECP signals. This platform could flexibly integrate devices for the intended purpose. In our recording system, there were a four-channel low-impedance headstage (RA4L1), a 16-channel preamplifier (RA16PA Medusa PreAmps), a DSP device (RA16BA Medusa Base Station), and a PC interface module (FI5/PI5-to-zBus). Signals were digitized at 25 kHz on the preamplifier and sent over a fiber optic link to a DSP device where they were filtered (0.5–300 Hz) and processed in real time. Ordinarily, 30 consecutive responses were summed and averaged on one VECP/EECP record.

For recording the VECP, a photopic stimulator (Neuropack 2 plus, Nihon Kohden, Tokyo, Japan) was positioned 3 cm above the rabbit eye. The stimulator intensity setting was 1.2J, and flash frequency was 1 Hz.

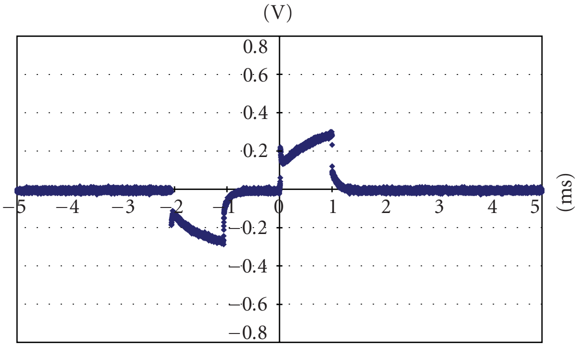


Figure 6: The voltage drop between a strip-shaped electrode (in the suprachoroidal space under visual streak) and the reference electrode (on the surface of sclera) with 104 uA current intensity (upon implantation.)

For recording the EECP, a cathodic-first biphasic constant current stimulus waveform was simultaneously applied to all selected active channels. Both pulse duration and interpulse delay were 1 millisecond and amplitudes were varied as noted. The repetition frequency was 4 Hz.

During the electrical stimulation, real-time signals also were recorded by a cornea contact electrode through another recording channel in the above recording system. From these recorded signals, the stimulation artifact signals were extracted and used as the trigger signal for EECP recording. In our measurements, the delay of the extracted trigger signal compared to the stimulation signal was less than 500 microseconds and had a constant value.

### 2.4. Long-term follow-up of rabbit retina and implanted electrodes

We observed the rabbits from 1 month to 16 months. We used OCT (optical coherence tomography, Stratus OCT Model 3000, Carl Zeiss Ophthalmic System, Dublin, Calif, USA) and fundus photography to evaluate the biocompatibility of the polyimide electrodes.

## 3. RESULTS

### 3.1. Retinal prosthesis system

A current mode, charge-balanced, cathodic-first biphasic stimulation waveform was generated in the stimulation mode and provided to a stimulation electrode array. The current stimulation chip could simultaneously deliver a stable current from 8 $\mu$A to 2 $\mu$A to all channels. The pulse width and the interphase delay could be changed up to 3 milliseconds. The interphase delay was designed to have the same time duration as the pulse width.

The fabricated polyimide electrode array was checked for electrode impedance in vitrousing a commercial potentiostat (Zahner Elektrik IM6e, Germany). Impedance for the strip-shaped electrode site (750 $\mu$m × 300 $\mu$m) was typically 1.3 kΩ and for the reference electrode was typically 300 Ω in phosphate-buffered solution (pH 7.4) as measured at 1 kHz.

Table 1: Measured impedance at 1 KHz in vivo for two weeks after implantation. The strip-shaped seven-segment electrode array was placed into the suprachoroidal space and reference electrode was placed on the surface of sclera. A strip-shaped electrode site was connected to the working electrode input, and the stimulating reference electrode served as the counter and reference electrode. 0 week means the day of implantation.

| Site No. | Impedence (kohm)/Phase (degree) | | |
|---|---|---|---|
| | 0 week | 1 week | 2 weeks |
| 1 | 10.50 (−42.8) | 9.50 (−44.5) | 9.99 (−45.7) |
| 2 | 14.10 (−40.3) | 13.30 (−35.0) | 13.00 (−29.3) |
| 3 | 9.01 (−43.2) | 9.48 (−41.6) | 11.23 (−39.7) |
| 4 | 13.70 (−41.3) | 10.40 (−38.9) | 11.80 (−39.4) |
| 5 | 9.29 (−38.0) | 10.94 (−40.9) | 9.08 (−36.9) |
| 6 | 10.25 (−45.9) | 9.67 (−43.0) | 8.98 (−40.7) |
| 7 | 8.94 (−43.8) | 10.08 (−46.7) | 12.60 (−39.2) |

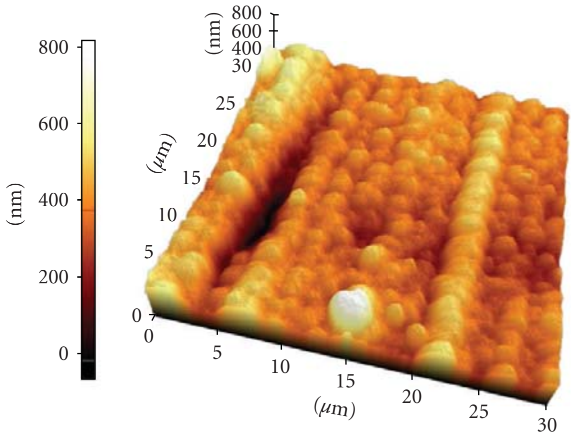


Figure 7: An AFM image of gold/polyimide electrode surface. The average surface roughness was 68 nm.

In vivo electrode impedance of the stimulation electrode site was checked for two weeks postsurgery. The potentiostat was connected to the sites through a percutaneous connector. Impedance between a strip-shaped site and the reference electrode was typically near 10 kΩ at 1 KHz and there was no significant change during the two-weeks monitoring period. The result is detailed in Table 1.

The surface of the gold stimulation site was checked with an atomic force microscopy (AFM) image (AFM, PSIA, XE-150) as shown in Figure 7. The surface average roughness of gold electrode array was 68 nm.

The stimulation outputs were connected to 1.3 kΩ resistive loads, modeling the electrodes. The stimulator consumed around 2 mW when delivering 520-$\mu$A biphasic current pulse with 1-millisecond pulse width at a stimulation rate of 4 Hz.

The capacity of the battery was 75 mAh (4.2 V) and the battery could supply the power to the internal circuitries for over 30 hours under the 520-$\mu$A amplitude, 1-millisecond stimulation duration. The battery was fully recharged within

Table 2: Specification summary of the implanted stimulator.

| | |
|---|---|
| Hermtically packaged stimulator | 35 mm(L) × 26 mm (W) × 8 mm (H) |
| Diameter of internal coil | 28 mm |
| Length of lead between electrode and stimulator | 20 cm |
| Stimulator electrode | 4 mm × 4 mm |
| Stimulation site of the electrode | 750 um × 300 um |
| Electrode site impedance | 1300 ± 106 ohm at PBS solution pH 7.4 |
| Number of current generator | 7 |
| Data packet | 22 bits |
| Stimulation duration (number of bits; minimum duration) | 4-bit (200 us) |
| Stimulation amplitude (number of bits; minimum amplitude) | 8-bit (8 uA) |
| Power consumption (520 uA, 1 ms, 4 Hz) | 2 mw |
| Rechargeable battery size | 20 mm (D) × 4 mm (H) |
| Rechargeable battery capacity | 75 mAh (4.2 V) |
| Recharging time | 3 hours (25 mA charging current) |

three hours with 25 mA charging current through the RF inductive link when in the battery recharging mode.

Table 2 summarizes the implantable stimulator specifications and performance.

### 3.2. Postoperative state of rabbits

Implantation of the stimulation system into the rabbit was successfully achieved. The internal portions were harbored safely and postoperatively remained in situ, as verified with fundus photography. There were no limitations in eye movements or shoulder movements on the implanted side and the entire stimulation system could be safely protected under the skin during the follow-up period. There was no need to anesthetize the rabbit while changing the stimulation parameters by attaching the external RF coil onto the skin overlaying the internal RF coil.

### 3.3. VECP & EECP

Figure 8 shows a typical VECP and EECP recording for a single rabbit with varying stimulus current. The overall shapes and latencies of VECP waves were similar to previous reports [19]. The EECP were well recorded upon electrical stimulation from the implanted electrodes. The EECP disappeared after the optic nerves were severed.

Threshold current necessary to elicit the EECP's were also recorded from another animal for six weeks post-surgery. Seven channels were stimulated simultaneously using a biphasic pulse repeated at 4 Hz with a 1-millisecond duration/phase and a 1-millisecond interphase delay. The threshold ranged from 80 $\mu$A to 128 $\mu$A during the testing period but did not show any significant change. The detailed measured threshold currents and calculated charge density values are shown in Table 3. The threshold was the highest (128 $\mu$A) on the day of implantation but then fell and remained at lower values for the duration of monitoring. The threshold charge density was similar to the value reported in [20].

Table 3: Threshold current to elicit the EECP. Seven channels were used simultaneously and the duration of the pulse was a biphasic current pulse, 1 ms/phase with a 1-millisecond interphase delay. The repetition rate was 4 Hz. 0 week means the day of implantation.

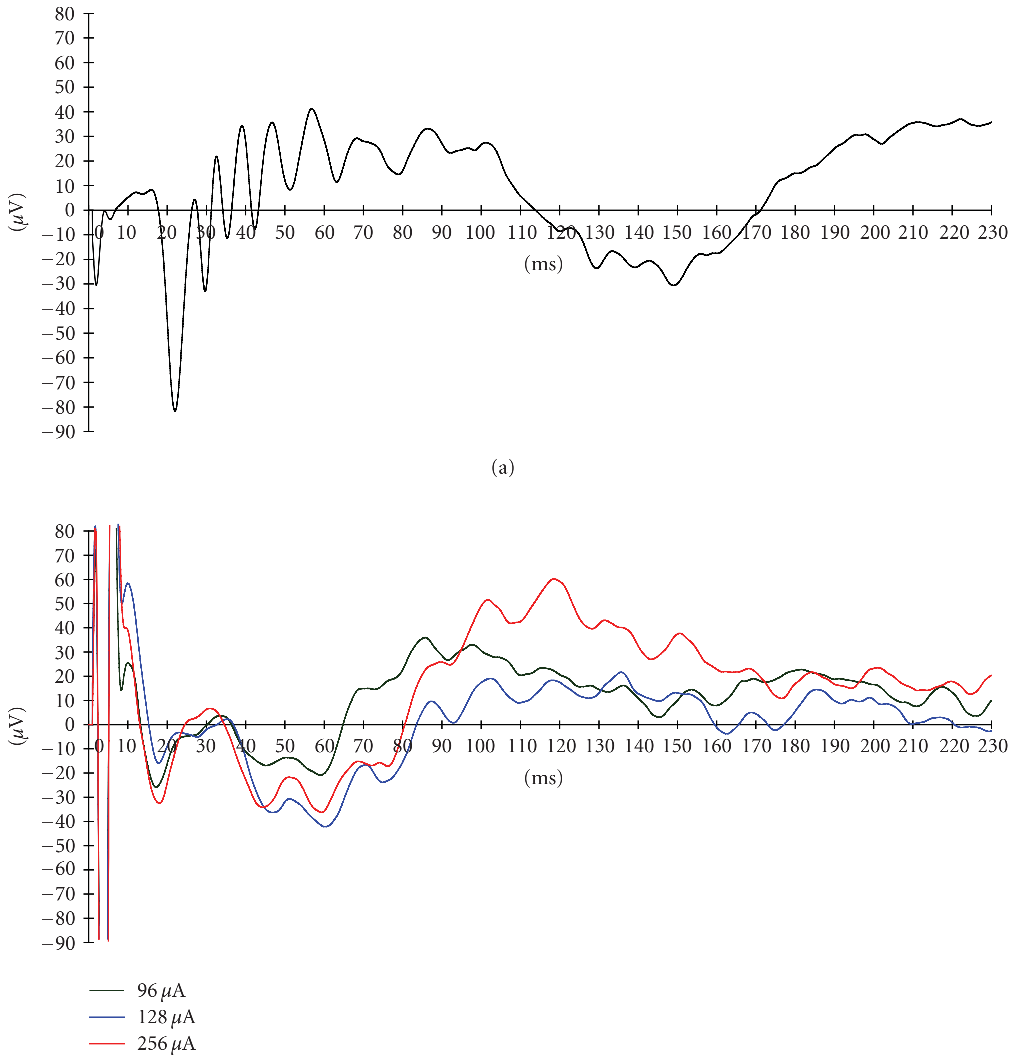
| Time after implantation | Threshold current ($\mu$A) | Threshold charge density per phase ($\mu$C/cm2) |
|---|---|---|
| 0 week | 128 | 48.64 |
| 1 week | 104 | 39.52 |
| 2 weeks | 80 | 30.40 |
| 3 weeks | 96 | 36.48 |
| 4 weeks | 80 | 30.40 |
| 5 weeks | 104 | 39.52 |
| 6 weeks | 80 | 30.40 |

### 3.4. Long-term follow-up results

The fundus photography and OCT findings are shown in Figure 9. The Stratus OCT 3000 showed fine resolution of cross-sectional retinal images of rabbits. We can distinguish each layer of retina and choroid by OCT. All the implanted electrodes were detected in the suprachoroidal space, which was in the same location of the retina as immediately after surgery. The eyes with polyimide-based gold electrodes showed no chorioretinal reaction around the electrodes (see Figure 9(a)). The implanted electrodes did not cause any chorioretinal inflammation and structural deformity during 16 months of the follow-up period.

## 4. DISCUSSION

In retinal prosthesis research, long-term electrical stimulation experiments in animal are needed to verify the long-term stability and durability of the system before clinical use. Retinal prosthesis systems for electrical stimulation usually consist of external and internal units [21, 22]. The external unit is necessary to change the stimulation parameters and to provide power for the implanted stimulator, but it is also



(a)

(b)

Figure 8: (a) Visual-evoked cortical potential (VECP) recorded before implantation of electrodes. (b) Electrically evoked cortical potentials (EECP) following retinal stimulation with a suprachoroidal electrode.

burdensome especially in long-term animal electrical stimulation experiments.

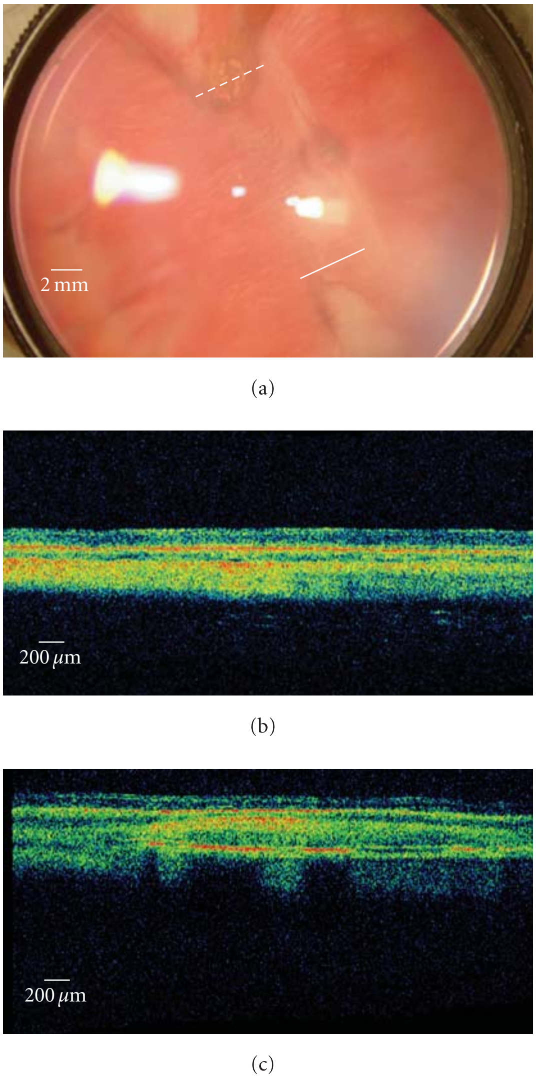
The stimulation system presented in this article was intended to provide a useful tool for long-term animal experiments on retinal prostheses. To remove the external connection or the external unit from the animal during electrical stimulation, we used a small rechargeable battery in the implantable stimulator. This battery could be simply recharged using an RF inductive link while the system was idle. This system makes it possible to conduct chronic electrical stimulation tests in such a way that the animal can move and act freely without any external unit restriction during the stimulation test. Therefore, there is no need to anesthetize the animal frequently and the stimulation system is also protected from the animal's claws and teeth.

The implantable retinal stimulator was built using IC chips and discrete elements for this proof-of-concept. An IC chip can be developed to reduce power consumption and further miniaturize the implanted component of the device.

During the implantation of the polyimide electrode array into the eye and the inside of orbit, we adopted Humayun's method [4], making one turn of the electrode along the equator of the eye under the extraocular muscles to provide stability. The elongated polyimide electrode is flexible, thus simple folding of the connection part can provide much degrees of freedom, and this enables eye movement without limitation of motion.

There are pros and cons regarding the ocular location of implantation of the electrodes. To our thought, the suprachoroidal implantation of electrodes has many advantages over the conventional subretinal implantation.

First, as the electrode does not contact retinal cells directly, there is no risk of mechanical retinal damage from direct contact with electrode which could be initiated



(a)

(b)

(c)

Figure 9: (a) A digital fundus photography of rabbit retina implanted with a suprachoroidal electrode with two OCT (Stratus OCT 3000) scan paths (solid and dashed lines). A polyimide-based gold electrode had been implanted. Under gross inspection, the electrode is in its original position near the visual streak and no chorioretinal abnormalities can be found. (b) The OCT scan of the retina remote from the electrode (the solid line in the fundus photography). The normal structures of the rabbit retina and choroid is demonstrated. (c) The OCT scan passing the electrode vertically (the dashed line in the fundus photography). The electrode is observed in the suprachoroidal space and in tight contact with the choroid. The metal plates in the electrode are highly reflective and leave posterior shadowing. The overlying retina showed normal structures of layers.

during surgical procedure. A past report comparing subretinal and suprachoroidal retinal implantation in rabbits also showed histologically proven retinal damage in the subretinal implantation while there was none in the suprachoroial implantation [12]. Second, in the suprachoroidal implantation method, the retina could be protected from heat damage generated from electrodes. As the choroidal blood flow is much larger than in other organs in human body, the large amount of choroidal blood flow will carry away the heat from the electrodes and protect the retina from heat damage [23]. Third, the surgical procedure of suprachoroidal implantation is more simple and safer than subretinal implantation. To introduce electrodes into the suprachoroidal space, it is not necessary to induce retinal tears and retinal detachment. In the case of retinal detachment during subretinal implantation, the electrode could inhibit retinal reattachement after implantation and could cause redetachement.

The suprachoroidal implantation also has weaknesses as compared to subretinal implantation. The threshold of suprachoroidal implantation was shown to be larger than the subretinal electrode [12]. The distance from the electrodes to inner retinal cells might influence the impedence of electric current and this may cause problems in making fine resolution pattern electrodes.

Our suprachoroidal approach was similar to the previous method [20] in many aspects. However, we placed the reference electrode on the outer scleral surface without penetrating vitreous cavity. In our result, the placement of the reference electrode in the extraocular space showed effective stimulation of the retina which was evidenced by EECP and was safer than the previous method [20] without risk of ocular damage from penetrating vitreous cavity.

In the present study, we confirmed the feasibility of our electrodes by observing the propagation of electrically induced signals to the visual cortex by a conventional recording method.

Moreover, the biocompatibility of the electrode was evaluated with OCT and fundus photography. The gold electrode was safe and biocompatible as late as 16 months after introduction. OCT gives in vivo cross-sectional retinal images of 10 $\mu$m resolution and is an in vivo biomicroscopy method [24, 25]. OCT was recently used to evaluate the subretinally implanted microphotodiode arrays in cats and pigs [26, 27] and epiretinal electrodes in dogs [8]. To our knowledge, we are the first to evaluate the biocompatibility of suprachoroidal electrodes which were implanted into rabbits. OCT was especially useful in detecting the subtle amount of retinal inflammation when no retinal pathology could be found with fundus photography.

In conclusion, we have demonstrated a new design for a retinal electrical stimulator which can be used safely in the long term. More animal experiments are needed to improve our system and to further develop an advanced system for use in humans.

## ACKNOWLEDGMENTS

This work was supported by the Korean Science and Engineering Foundation (KOSEF) through the Nano Bioelectronics and Systems Research Center (NBS-ERC) of Seoul National University under Grant No. R11-2000-075-01001-0 and by the Ministry of Health & Welfare (South Korea) through the Nano Artificial Vision Research Center under Grant of the Korea Health 21 R&D Project (A050251). This material was presented, in part, at The World Congress on Engineering (ICSBB 2007), July 2007.

## REFERENCES

[1] I. Wickelgren, "A vision for the blind," *Science*, vol. 312, no. 5777, pp. 1124–1126, 2006.

[2] P. C. Hessburg and J. F. Rizzo III, "The eye and the chip. World congress on artificial vision 2006," *Journal of Neural Engineering*, vol. 4, no. 1, 2007.

[3] J. F. Rizzo III, J. Wyatt, J. Loewenstein, S. Kelly, and D. Shire, "Perceptual efficacy of electrical stimulation of human retina with a microelectrode array during short-term surgical trials," *Investigative Ophthalmology & Visual Science*, vol. 44, no. 12, pp. 5362–5369, 2003.

[4] M. S. Humayun, J. D. Weiland, G. Y. Fujii, et al., "Visual perception in a blind subject with a chronic microelectronic retinal prosthesis," *Vision Research*, vol. 43, no. 24, pp. 2573–2581, 2003.

[5] R. Hornig, T. Laube, P. Walter, et al., "A method and technical equipment for an acute human trial to evaluate retinal implant technology," *Journal of Neural Engineering*, vol. 2, no. 1, pp. S129–S134, 2005.

[6] A. Y. Chow, V. Y. Chow, K. H. Packo, J. S. Pollack, G. A. Peyman, and R. Schuchard, "The artificial silicon retina microchip for the treatment of vision loss from reinitis pigmentosa," *Archives of Ophthalmology*, vol. 122, no. 4, pp. 460–469, 2004.

[7] A. Y. Chow, M. T. Pardue, J. I. Perlman, et al., "Subretinal implantation of semiconductor-based photodiodes: durability of novel implant designs," *Journal of Rehabilitation Research and Development*, vol. 39, no. 3, pp. 313–322, 2002.

[8] D. Güven, J. D. Weiland, M. Maghribi, et al., "Implantation of an inactive epiretinal poly(dimethyl siloxane) electrode array in dogs," *Experimental Eye Research*, vol. 82, no. 1, pp. 81–90, 2006.

[9] T. Schanze, H. G. Sachs, C. Wiesenack, U. Brunner, and H. Sailer, "Implantation and testing of subretinal film electrodes in domestic pigs," *Experimental Eye Research*, vol. 82, no. 2, pp. 332–340, 2006.

[10] J. M. Seo, S. J. Kim, H. Chung, E. T. Kim, H. G. Yu, and Y. S. Yu, "Biocompatibility of polyimide microelectrode array for retinal stimulation," *Materials Science and Engineering C*, vol. 24, no. 1-2, pp. 185–189, 2004.

[11] P. Walter and K. Heimann, "Evoked cortical potentials after electrical stimulation of the inner retina in rabbits," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 238, no. 4, pp. 315–318, 2000.

[12] Y. Yamauchi, L. M. Franco, D. J. Jackson, et al., "Comparison of electrically evoked cortical potential thresholds generated with subretinal or suprachoroidal placement of a microelectrode array in the rabbit," *Journal of Neural Engineering*, vol. 2, no. 1, pp. S48–S56, 2005.

[13] E. Margalit, J. D. Weiland, R. E. Clatterbuck, et al., "Visual and electrical evoked response recorded from subdural electrodes implanted above the visual cortex in normal dogs under two methods of anesthesia," *Journal of Neuroscience Methods*, vol. 123, no. 2, pp. 129–137, 2003.

[14] E. Zrenner, D. Besch, K. U. Bartz-Schmidt, et al., "Subretinal chronic multi-electrode arrays implanted in blind patients," *Invesigative Ophthalmology & Visual Science*, vol. 47, p. 1538, 2006.

[15] M. S. Humayun, J. Hopkins, S. H. Greenwald, et al., "Electrical effects and perceptual performance using a chronically implanted 16-channel epiretinal prosthesis in blind subjects," *Invesigative Ophthalmology & Visual Science*, vol. 47, p. 3212, 2006.

[16] D. Güven, J. D. Weiland, G. Fujii, et al., "Long-term stimulation by active epiretinal implants in normal and RCD1 dogs," *Journal of Neural Engineering*, vol. 2, no. 1, pp. S65–S73, 2005.

[17] J. A. Zhou, E. T. Kim, J. M. Seo, H. Chung, and S. J. Kim, "A seven segment electrode stimulation system for retinal prosthesis," *Invesigative Ophthalmology & Visual Science*, vol. 47, p. 3178, 2006.

[18] S. K. An, S. I. Park, S. B. Jun, et al., "Design for a simplified cochlear implant system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 6, pp. 973–982, 2007.

[19] T. Okuno, H. Oku, and T. Ikeda, "The reproducibility and sensitivity of visual evoked potential testing in rabbits," *Neuro-Ophthalmology*, vol. 26, no. 1, pp. 59–66, 2001.

[20] H. Sakaguchi, T. Fujikado, X. Fang, et al., "Transretinal electrical stimulation with a suprachoroidal multichannel electrode in rabbit eyes," *Japanese Journal of Ophthalmology*, vol. 48, no. 3, pp. 256–261, 2004.

[21] G. J. Suaning and N. H. Lovell, "CMOS neurostimulation ASIC with 100 channels, scaleable output, and bidirectional radio-frequency telemetry," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 2, pp. 248–260, 2001.

[22] W. Liu, M. Sivaprakasam, P. R. Singh, R. Bashirullah, and G. Wang, "Electronic visual prosthesis," *Artificial Organs*, vol. 27, no. 11, pp. 986–995, 2003.

[23] L. M. Parver, C. Auker, and D. O. Carpenter, "Choroidal blood flow as a heat dissipating mechanism in the macula," *American Journal of Ophthalmology*, vol. 89, no. 5, pp. 641–646, 1980.

[24] C. A. Puliafito, M. R. Hee, C. P. Lin, et al., "Imaging of macular diseases with optical coherence tomography," *Ophthalmology*, vol. 102, no. 2, pp. 217–229, 1995.

[25] J. G. Fujimoto, D. Huang, M. R. Hee, et al., "Physical principles of optical coherence tomography," in *Optical Coherence Tomography of Ocular Diseases*, J. S. Schuman, C. A. Puliafito, and J. G. Fujimoto, Eds., pp. 677–698, SLACK, Thorofare, NJ, USA, 2nd edition, 2004.

[26] M. Völker, K. Shinoda, H. G. Sachs, et al., "In vivo assessment of subretinally implanted microphotodiode arrays in cats by optical coherence tomography and fluorescein angiography," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 242, no. 9, pp. 792–799, 2004.

[27] F. Gekeler, P. Szurman, S. Grisanti, et al., "Compound subretinal prostheses with extra-ocular parts designed for human trials: successful long-term implantation in pigs," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 245, no. 2, pp. 230–241, 2007.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 681303, 8 pages
doi:10.1155/2008/681303

*Research Article*

# Multimodal Data Integration for Computer-Aided Ablation of Atrial Fibrillation

**Jonghye Woo,[1] Byung-Woo Hong,[2] Sunil Kumar,[3] Indranill Basu Ray,[4] and C.-C. Jay Kuo[1]**

[1] *Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564, USA*
[2] *School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Korea*
[3] *Electrical and Computer Engineering Department, San Diego State University, San Diego, CA 92182, USA*
[4] *School of Medicine and Dental Sciences, State University of New York at Buffalo, Buffalo, NY 14214, USA*

Correspondence should be addressed to Byung-Woo Hong, hong@cs.ucla.edu

Received 30 August 2007; Accepted 17 December 2007

Recommended by Daniel Howard

Image-guided percutaneous interventions have successfully replaced invasive surgical methods in some cardiologic practice, where the use of 3D-reconstructed cardiac images, generated by magnetic resonance imaging (MRI) and computed tomography (CT), plays an important role. To conduct computer-aided catheter ablation of atrial fibrillation accurately, multimodal information integration with electroanatomic mapping (EAM) data and MRI/CT images is considered in this work. Specifically, we propose a variational formulation for surface reconstruction and incorporate the prior shape knowledge, which results in a level set method. The proposed method enables simultaneous reconstruction and registration under nonrigid deformation. Promising experimental results show the potential of the proposed approach.

## 1. INTRODUCTION

Current treatment of cardiac arrhythmias ranges from noninvasive strategies, such as pharmacological therapy, to minimally invasive techniques, such as catheter-based ablation, and to open surgical techniques. While medical therapy can mitigate the occurrence of arrhythmias, these treatments may have significant side effects since most drugs used have some toxicity that is not suitable for long-term therapy. The catheter-based procedure is proven to be an effective method in treating patients with certain cardiac arrhythmias [1]. It is much less invasive and more established. It also demands shorter recovery time than the surgical approach. Thus, catheter-based radio frequency (RF) ablation has become a widely accepted method in the treatment of cardiac arrhythmias, including atrial fibrillation (AF) and ventricular tachycardia (VT). These arrhythmias affect a large number of people and result in significant morbidity and mortality.

AF is the most common sustained cardiac arrhythmia encountered in clinical practice. In the United States alone, there are over 3.5 million patients with this disorder [2]. AF can result in serious complications, including congestive heart failure and thromboembolism. Despite recent advances, drug therapy to control this disease is still unsatisfactory. As an alternative, a nonpharmacological, interventional approach based on creating percutaneous catheter-based lesion inside the heart has been developed. Lesions are delivered in the left atrial-pulmonary vein junction with an aim to electrically isolate these veins from the rest of the atrium. This protects the atrium from fast heart beating that is originated in the veins, which initiate and perpetuate AF.

The procedure of interventional AF treatment entails mapping the left atrium and the attached pulmonary veins using an electroanatomic mapping (EAM) system. This mapping information can be used to deliver lesions as well. This electrical approach is suitable for the heart since it is an electromechanical organ, where mechanical contractions are driven by electrical stimulus. However, there is a serious limitation of the EAM system in that it is not able to provide an accurate anatomical information of heart. Typically, a virtual shell is used to represent the atrial wall and the vein. The points on the atrial wall, where the catheter is manually touched, are used to create this shell [3].

The catheter-based ablation process can be greatly improved if a real anatomy is used instead of the virtual shell. To ensure safe catheter maneuverability and enable delivery of effective lesions with minimal collateral damage and complications, it is critical to have both the anatomical information and the electrical information available to the operator. This is particularly important for performing ablation in a complex structure such as the left atrium that is surrounded by important organs, which are vulnerable to damage if lesions are not appropriately directed with close anatomical guidance. Furthermore, even the pulmonary veins themselves are liable to be damaged with grave long-term consequences if the lesions extend deeply into the veins instead of being restricted to the ostia.

The use of a multimodal data integration process can provide an anatomical, physiological, and functional representation simultaneously. In practice, this can be achieved by combining an anatomical surface model acquired by MRI/CT images and the localized electrical information measured by an EAM system. In the registration process, one obvious difficulty stems from the noise and/or outliers that are inevitably associated with the MRI/CT imaging process and the EAM data collection procedure. Unlike other organs in the body, heart undergoes contractile motion, apart from respiratory motion, thus making it unique and very challenging to register and integrate data of different modalities. In addition to physiological variations such as changes in the heart rate, the heart rhythm, and the respiratory effect, various types of heart motion are the source of outliers.

The main contribution of this work is to provide enhanced imaging of the anatomical heart surface from sparse and noisy EAM data in combination with a heart-shape model obtained from MRI/CT reconstruction as a prior knowledge. For 3D surface reconstruction, we propose a variational formulation in the level set framework that is an efficient numerical scheme. The level set method is particularly of great use in representing a shape due to its topology-free and implicit characteristics. By leveraging the 3D heart-shape model, we can compensate incomplete EAM data, thereby representing the anatomical heart more accurately. The proposed method has two important advantages. First, it is robust against nonrigid deformation caused by cardiac motion and noise. Second, it can construct the optimal surface without an explicit correspondence between the MRI/CT surface and EAM data due to the implicit surface representation.

The rest of this paper is organized as follows. Previous related work is reviewed in Section 2 followed by the proposed multimodal data integration method for computer-aided ablation of atrial fibrillation in Section 3. Both synthetic and real data are tested to demonstrate the efficiency of the proposed method in Section 4. Concluding remarks and future work are given in Section 5.

## 2. REVIEW OF RELATED WORK

Developing a computer-guided system for ablative heart surgery involves image registration or integration techniques. They are usually performed under a rigid transformation between preoperative MRI/CT reconstruction and intraoperative EAM data points [4, 5]. Among various registration algorithms, the iterative closest point (ICP) method and its variants have been widely used for this application due to their computational efficiency [6].

The ICP algorithm begins with two meshes and an initial guess for their relative rigid-body transform. It refines the transform iteratively by generating pairs of corresponding points on the meshes and minimizing an error metric repeatedly [7]. However, the standard ICP algorithm does not take noise and outliers into account. Since noise and outliers may affect the ICP performance substantially, several ICP variants have been proposed in [8] to mitigate this problem. One popular approach to identify outliers is to use a threshold, including a certain constant, a fraction of a sorted distance, and some multiple of the standard deviation of a distance [9–11]. Even with these variants, it is still challenging to deal with nonrigid deformation and differentiate inliers from outliers.

Most of previous schemes used an ICP-based method without addressing the above-mentioned problem. Instead, they focused on clinical registration. For example, Reddy et al. [5] and Malchano [6] showed the feasibility of combining MRI with CARTO-XP in a porcine model of myocardial infarction (MI). They used the modified Iterative Closest Point (mICP) scheme for registration, but did not address the outlier problem. The modification is to adopt hierarchical registration by adding the class information in the algorithm. A clinical registration strategy that combines landmarks and surface registration was proposed in [4]. This study assessed the accuracy for each cardiac chamber using a different clinical registration method. It was observed in [12] that the size of the left atrium affects the accuracy. The patient who has a bigger chamber volume tends to have more ablation errors.

The rigid transformation assumption made by existing schemes is simple yet insufficient in most cases. It often yields unsatisfactory results since a nonrigid deformation is involved between the anatomical heart model reconstructed by MRI/CT images and temporal instances of the heart at the collection of EAM data points. This physiological and anatomical variation that occurs in the formation of the heart surface model and the collection of EAM data points demands a nonrigid transformation (or equivalently diffeomorphism) between the model and the data. Woo et al. proposed a novel image integration technique by incorporating nonrigid deformation using the level set method in [13].

To overcome the limitation of the traditional registration approach based on the rigid-transformation assumption, we formulate this problem as a 3D surface reconstruction problem from EAM data points with a given surface prior. A similar context arises in surface reconstruction from point clouds in a scanned noisy image. Surface reconstruction using an explicit representation has been considered by researchers, for example, [14, 15]. Typically, this approach needs to parameterize a large point set that could be difficult to manipulate. Another approach was proposed in [16, 17] to construct triangulated surfaces using Delaunay triangulations and Voronoi diagrams. It has to determine the right connection among points in the point set, which could be challenging in handling noisy and unorganized point data.

Surface reconstruction based on an implicit shape representation using the level set technique has been studied for almost two decades by applied mathematicians, for example, [18–20]. The nonparametric (or implicit) surface representation has an advantage in dealing with arbitrary topology change and deformation. Hoppe et al. [21] proposed an algorithm in reconstructing a surface using the signed distance function from unorganized points. Zhao et al. [22] proposed another algorithm using the unsigned distance function and the weighted minimal energy to reconstruct the surface. These algorithms are however restricted to situations where the population of points is dense enough to characterize the target surface. They are not applicable to our application where EAM data points are sparse and insufficient.

The problem of insufficient EAM data encountered in the 3D heart surface construction, including the left atrium and its pulmonary veins, can be mitigated by incorporating a heart-shape prior provided by MRI/CT imaging. Then, the optimal surface can be obtained by minimizing the energy functional that consists of a data-fitting term and a prior knowledge term as detailed in the next section.

## 3. MULTIMODAL DATA INTEGRATION FOR SIMULTANEOUS SURFACE RECONSTRUCTION AND REGISTRATION

In this section, we present a multimodal data integration algorithm for simultaneous surface reconstruction and registration. This algorithm reconstructs the heart surface from measured EAM data points and a heart-shape prior obtained by MRI/CT imaging using the level set method.

### 3.1. Surface reconstruction and registration

Under the level set framework [23], a surface $S(x)$ in $\mathbb{R}^3$ (a curve in $\mathbb{R}^2$) to reconstruct can be represented by the zero-level set of a higher dimensional embedding function $\phi(x) : \Omega \rightarrow \mathbb{R}$ as given by

$$\begin{aligned} S &= \{x \in \Omega \mid \phi(x) = 0\}, \\ \text{interior}(S) &= \{x \in \Omega \mid \phi(x) > 0\}, \\ \text{exterior}(S) &= \{x \in \Omega \mid \phi(x) < 0\}, \end{aligned} \tag{1}$$

where $\Omega$ is the domain of $\phi(x)$. This implicit representation can provide the geometric shape of surface with the region defined by the union of $S$ and its interior, denoted by $\overline{S}$. Then, the shape of $S$ is given by $\overline{S}(x) = H(\phi(x))$, where $H(x)$ is the heaviside function as defined by

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2}$$

For the representation of a surface model $M(x)$ obtained from MRI/CT images, we define its shape $\overline{M}(x)$ in a similar way using another level set function $psi(x)$ as given by $H(\psi(x))$. Now, we denote the set of measured EAM data points by

$$D = \{p_1, p_2, \ldots, p_n\} \subset \mathbb{R}^3, \tag{3}$$

where $n$ is the number of data points.

The essential assumption in this surface reconstruction application is that surface $S$ to reconstruct is an equivalent class to a given prior surface model $M$ under small nonrigid deformation and the surface is close to the data points in $D$. Thus, surface $S$ can be obtained by minimizing an energy functional that consists of a data fitting term and a prior knowledge term. Our goal is to find the embedding function $\phi$ associated with surface $S$ that minimizes the following energy functional:

$$E(\phi) = E_{\text{point}}(\phi, D) + \alpha E_{\text{prior}}(\phi), \tag{4}$$

where detail of each term will be described in the following section.

### 3.2. Derivation of energy terms

The energy functional in (4) consists of two terms. The first term measures how well the surface is fit to measured points based on the distance between the surface and these points. The second term measures how plausible the surface is in terms of the prior knowledge of the target surface. Parameter $\alpha \geq 0$ is a weight that adjusts the importance of these two factors.

The data fitting term can be written as

$$E_{\text{point}}(\phi | D) = \sum_{i=1}^{n} \int_{\Omega} |\phi(x) \cdot \delta(x - p_i)|^2 dx, \tag{5}$$

where $\delta(x) = (d/dx)H(x)$ is Dirac measure and this term sums up the Euclidean distance between point $p_i$ and $\phi$. Recall that $\phi$ is a signed distance function where the value of each point gives the euclidean distance between the point and the interface. To impose the prior knowledge on the target surface, $S$ should be close to prior surface model $M$ with a smooth-surface assumption. In other words, we penalize any abrupt change of the surface gradient. Here, we use two terms to represent the prior knowledge, that is,

$$E_{\text{prior}}(\phi) = E_{\text{reg}}(\phi) + E_{\text{shape}}(\phi \mid \psi), \tag{6}$$

where $E_{\text{reg}}(\phi)$ is the smoothness regularization term in the form of

$$E_{\text{reg}}(\phi) = \int_{\Omega} |\nabla H(\phi(x))| dx, \tag{7}$$

and $E_{\text{shape}}(\phi \mid \psi)$ is the shape dissimilarity term of the following form:

$$E_{\text{shape}}(\phi \mid \psi) = \int_{\Omega} |H(\phi(x)) - H(\psi(T(x)))|^2 dx, \tag{8}$$

and where $T(x)$ is a rigid transformation resulting from scaling, rotation, and translation. Note that the smoothness regularization term measures the length of the 2D curve and the area of a 3D surface while the shape dissimilarity term measures the symmetric difference between $H(\phi)$ and $H(\psi)$ under a rigid transformation.

Combining (4)–(8), the total energy functional $E(\phi)$ is given by

$$E(\phi) = E_{\text{point}}(\phi \mid D) + \alpha\left(E_{\text{reg}}(\phi) + E_{\text{shape}}(\phi \mid \psi)\right). \tag{9}$$

Then, surface reconstruction can be achieved using the energy minimization principle as

$$\phi^{*} = \arg\min_{\phi} E(\phi) \tag{10}$$

under constraint $|\nabla\phi| = 1$, which is a property of a signed distance function. Instead of applying the same weight $\alpha$ for both $E_{\text{reg}}$ and $E_{\text{shape}}$ as shown in (9), different weights can be used for each term.

### 3.3. Numerical implementation

For numerical implementation, we use the following approximations for the heaviside function and the Dirac delta measure as defined in [24, 25]:

$$\delta(z) = \begin{cases} 0 & \text{if } |z| > \varepsilon, \\ \dfrac{1}{2\varepsilon}\left[1 + \cos\left(\dfrac{\pi z}{\varepsilon}\right)\right] & \text{if } |z| \le \varepsilon, \end{cases}$$
$$H(z) = \begin{cases} 1 & \text{if } z > \varepsilon, \\ 0 & \text{if } z < -\varepsilon, \\ \dfrac{1}{2}\left[1 + \dfrac{z}{\varepsilon} + \dfrac{1}{\pi}\sin\left(\dfrac{\pi z}{\varepsilon}\right)\right] & \text{if } |z| \le \varepsilon, \end{cases} \tag{11}$$

The energy functional in (10) can be minimized with respect to $\phi(x)$ using the Euler-Lagrange equation with $\phi(t = 0, x) = \phi_0(x)$ defining the initial surface. Finally, the gradient descent method is applied to the resultant Euler-Lagrange equation, which leads to

$$\begin{aligned} \frac{\partial\phi}{\partial t} &= -2\sum_{i=1}^{n}\phi(x)\delta(x - p_i) \\ &\quad + \alpha\left[\delta(\phi)\operatorname{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right) - 2\int_{\Omega}(H(\phi) - H(\psi))\delta(\phi)dx\right], \\ \phi(t = 0, x) &= \phi_0(x) \qquad \text{in } \Omega, \\ \frac{\delta(\phi)}{|\nabla\phi|}\frac{\partial\phi}{\partial\vec{n}} &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{12}$$

where $\vec{n}$ denotes the exterior normal to the boundary $\partial\Omega$, and $\partial\Omega/\partial\vec{n}$ denotes the normal derivative of $\phi$ at the boundary.

To discretize the equation in $\phi$, we adopt a finite differences explicit scheme. The usual notations are as follows: let $h$ be the space step and let $\Delta t$ be the time step. The finite differences are expressed as

$$\begin{aligned} \Delta^{x}\phi &= \frac{\phi_{i+1,j,k} - \phi_{i-1,j,k}}{2\Delta x}, \\ \Delta^{y}\phi &= \frac{\phi_{i,j+1,k} - \phi_{i,j-1,k}}{2\Delta y}, \\ \Delta^{z}\phi &= \frac{\phi_{i,j,k+1} - \phi_{i,j,k-1}}{2\Delta z}, \\ \Delta^{xx}\phi &= \frac{\phi_{i+1,j,k} + \phi_{i-1,j,k} - \phi_{i,j,k}}{\Delta x^{2}}, \\ \Delta^{yy}\phi &= \frac{\phi_{i,j+1,k} + \phi_{i,j-1,k} - \phi_{i,j,k}}{\Delta y^{2}}, \\ \Delta^{zz}\phi &= \frac{\phi_{i,j,k+1} + \phi_{i,j,k-1} - \phi_{i,j,k}}{\Delta z^{2}}, \\ \Delta^{xy}\phi &= \frac{\phi_{i+1,j+1,k} - \phi_{i+1,j-1,k} - \phi_{i-1,j+1,k} + \phi_{i-1,j-1,k}}{4\Delta x\Delta y}, \\ \Delta^{xz}\phi &= \frac{\phi_{i+1,j,k+1} - \phi_{i+1,j,k-1} - \phi_{i-1,j,k+1} + \phi_{i-1,j,k-1}}{4\Delta x\Delta z}, \\ \Delta^{yz}\phi &= \frac{\phi_{i,j+1,k+1} - \phi_{i,j+1,k-1} - \phi_{i,j-1,k+1} + \phi_{i,j-1,k-1}}{4\Delta y\Delta z}. \end{aligned} \tag{13}$$

We first set $\phi^{n}$ as the initial surface and then update $\phi^{n+1}$ using the following discretization:

$$\begin{aligned} \frac{\phi^{n+1}_{i,j,k} - \phi^{n}_{i,j,k}}{\Delta t} &= -2\sum_{m=1}^{n}\phi^{n}_{i,j,k}\delta(x - p_m) + \frac{\alpha}{h^{2}}\delta\left(\phi^{n}_{i,j,k}\right)\kappa \\ &\quad - 2\int_{\Omega}\left(H\left(\phi^{n}_{i,j,k}\right) - H(\psi)\right)\delta\left(\phi^{n}_{i,j,k}\right)dx, \end{aligned} \tag{14}$$

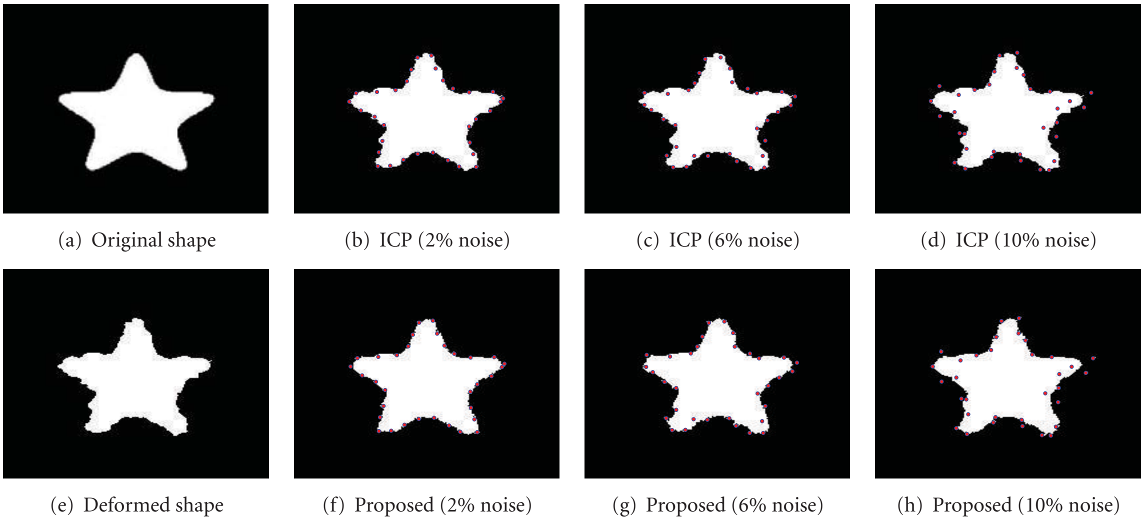
where

$$\begin{aligned} \kappa = \big(&\Delta^{x}\phi^{2}\Delta^{yy}\phi - 2\Delta^{x}\phi\Delta^{y}\phi\Delta^{xy}\phi + \Delta^{y}\phi^{2}\Delta^{xx}\phi \\ &+ \Delta^{x}\phi^{2}\Delta^{zz}\phi - 2\Delta^{x}\phi\Delta^{z}\phi\Delta^{xz}\phi + \Delta^{z}\phi^{2}\Delta^{xx}\phi \\ &+ \Delta^{y}\phi^{2}\Delta^{zz}\phi - 2\Delta^{y}\phi\Delta^{z}\phi\Delta^{yz}\phi\big). \end{aligned} \tag{15}$$

Solving the above partial differential equations numerically is challenging since the time step should be constrained to a small value in maintaining numerical stability. In addition, it is computationally expensive to find a high dimensional surface. Thus, it is desired to employ an efficient numerical scheme and we naturally use multigrid method that adopts a hierarchical representation of the data in multiple scales and propagates the solution from the coarse scale to the fine scale to achieve computational efficiency.

### 3.4. Relationship between variational formulation and Bayesian inference

This variational approach presented in Sections 3.1 and 3.2 can be interpreted from the interpretation of Bayesian inference under a probabilistic framework. This relationship is


(a) Original shape (b) ICP (2% noise) (c) ICP (6% noise) (d) ICP (10% noise)

(e) Deformed shape (f) Proposed (2% noise) (g) Proposed (6% noise) (h) Proposed (10% noise)

Figure 1: The shape reconstruction results for a synthetic 2D star shape, (a) the original shape, (b)–(d) ICP results using different Gaussian noise levels, (e) the deformed shape, and (f)–(h) results of the proposed method using different Gaussian noise levels.

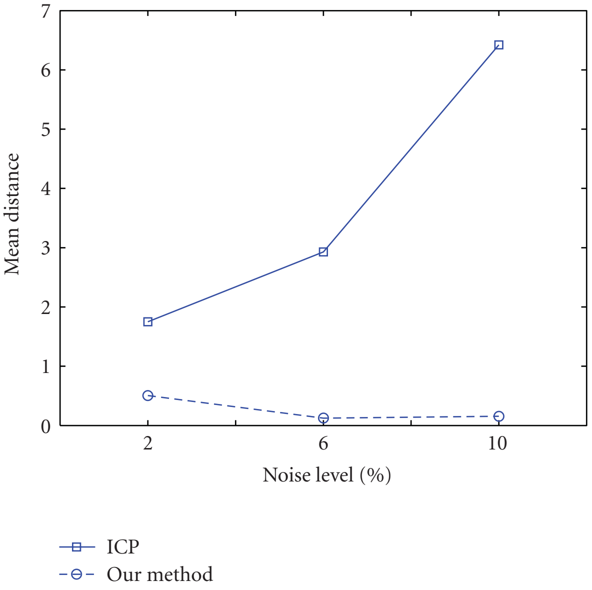


Figure 2: Comparison of mean distances between the reconstructed surface and measured data points for the 2D star example at different noise levels using ICP and the proposed algorithm.

presented in this subsection. The target surface $S$ can be obtained by maximizing the following posterior probability:

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)}, \tag{16}$$

where $P(D \mid S)$ is the likelihood function, $P(S)$ is the prior probability of the surface. Maximizing this conditional probability with data point $D$ for surface $S$ is equivalent to minimizing its negative logarithm

$$-\log\left(P(S \mid D)\right) = -\log\left(P(D \mid S)\right) - \log\left(P(S)\right) + c, \tag{17}$$

where $c$ is a constant. Thus, by setting

$$\begin{aligned} E_{\text{point}}(\phi, D) &= -\log\left(P(D \mid S)\right), \\ \alpha E_{\text{prior}}(\phi) &= -\log\left(P(S)\right), \end{aligned} \tag{18}$$

we can convert (17) to (4).

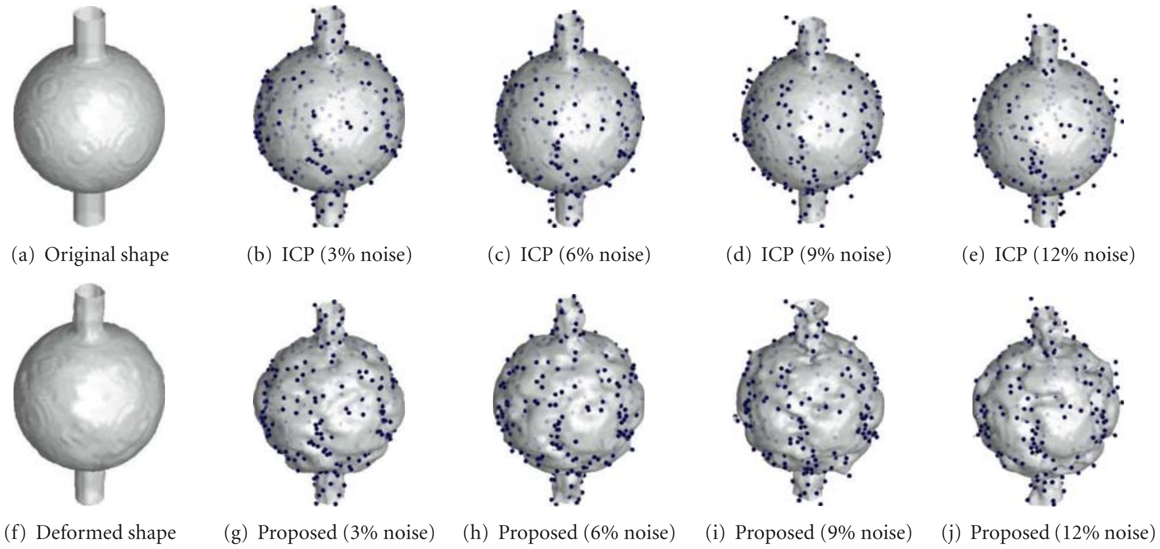
## 4. RESULTS AND DISCUSSION

We begin with simple yet illustrative examples to demonstrate the efficiency and robustness of the proposed algorithm. We use synthetic geometric objects in 2D and 3D which have geometric features that aim to be preserved under reconstruction process. Then, the evaluation of the algorithm is performed based on real patient data.

### *4.1. Synthetic data*

We first compare the proposed scheme with the ICP scheme in registration accuracy using a synthetic 2D star shape as shown in Figure 1. The original 2D star shape (image size: 200 × 200 pixels) is shown in Figure 1(a). It is deformed as shown in Figure 1(e). Furthermore, 33 noisy contour points are generated by adding Gaussian noise (2%, 6%, and 10% standard deviation of contour points, resp.) to original points obtained by sampling the original shape in Figure 1(a). The deformed shape stands for the reconstructed heart shape and data points with different noise levels represent the EAM data points. EAM data points can have errors from sensor, heart movement, and patient breathing. In synthetic experiments, we used Gaussian noise to represent noises introduced in EAM data points.

Graphical illustration of the registration results between deformed shape of different noise level and noisy contour points are presented in Figures 1(b)–1(d) for the ICP scheme. In Figures 1(f)–1(h), shape reconstruction using both shape prior Figure 1(e) and noisy contour points are presented



(a) Original shape (b) ICP (3% noise) (c) ICP (6% noise) (d) ICP (9% noise) (e) ICP (12% noise)

(f) Deformed shape (g) Proposed (3% noise) (h) Proposed (6% noise) (i) Proposed (9% noise) (j) Proposed (12% noise)

FIGURE 3: The shape reconstruction results for a synthetic 3D image: (a) the original shape, (b)–(e) ICP results using different Gaussian noise levels, (f) the deformed shape, and (g)–(j) results of the proposed method using different Gaussian noise levels.

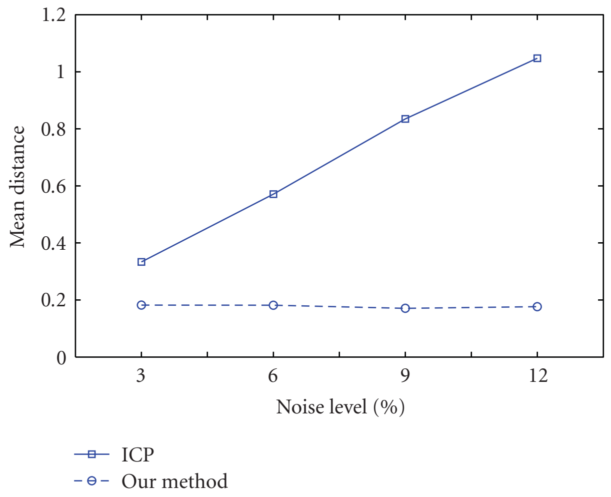
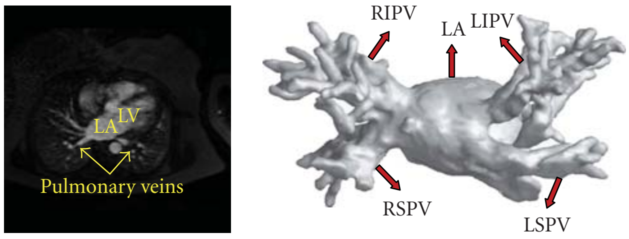


FIGURE 4: Comparison of mean distances between the reconstructed surface and measured data points for the 3D jar example at different noise levels using ICP and the proposed algorithm.



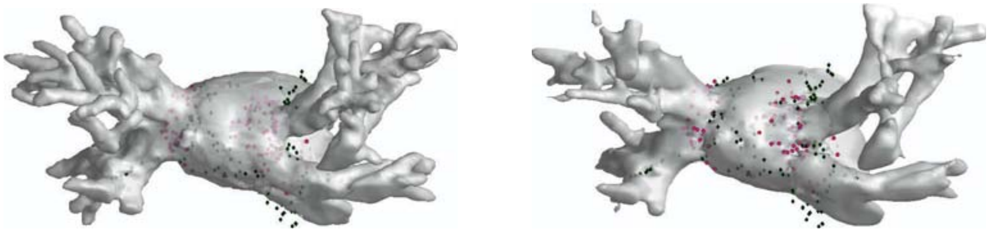
(a) MRA of LA (b) 3D Reconstruction

FIGURE 5: The 3D patient data: (a) MRA of LV and (b) 3D reconstruction result of the given MRA.

using the proposed scheme. It is clear from Figure 2 that the proposed scheme outperforms ICP. For quantitative error analysis, we measure the mean Euclidean distance between the reconstructed surface and measured data points with varying degree of Gaussian noise. The result is shown in Figure 2. Again, the proposed algorithm outperforms ICP significantly. This is especially true when the noise level is higher. The proposed algorithm produces more stable mean Euclidean distance than ICP as well.

Next, we compare the proposed scheme with ICP using a 3D synthetic jar example. Experimental results are shown in Figure 3, where the original and the deformed shapes are shown in Figures 3(a) and 3(f), respectively. Points extracted from the corrupted surfaces with various Gaussian noise levels (3%, 6%, 9%, and 12%) are used for visual evaluation. Registration between deformed surface and data points with different noise levels using ICP is shown in Figures 3(b)–3(e). Reconstructed surfaces based on data points at different noise levels with the proposed algorithm are presented in Figures 3(g)–3(j). The proposed method generated the reconstructed surface which incorporates the data point as well as deformed prior shape in Figure 3(e). The mean distances are also measured for accuracy comparison as shown in Figure 4. Again, the proposed algorithm is significantly better than the ICP scheme in terms of stability and distance, which is especially obvious at higher noise levels, as shown in Figure 4.

We can adjust the importance of both a data fitting term and a prior knowledge term that includes a regularization and shape similarity term by tuning the parameter $\alpha$. For the choice of the parameter $\alpha$, we use $\alpha = 1/3$ that gives reasonable results. In general, if $\alpha$ is too small, then the importance of the data fitting term increases, which results in overfitting to the data point. On the other hand, too big $\alpha$ produces the reconstructed shape which is almost same as the prior shape. Thus, parameter is desired to be carefully chosen according to the application.



(a) ICP

(b) Proposed method

Figure 6: The surface registration results for the patient data.

Table 1: Performance comparison of ICP and the proposed method.

| | ICP | Proposed |
|---|---|---|
| EAM point mean distance | 4.5087 mm | 2.4113 mm |
| Ablation point mean distance | 3.2046 mm | 2.0921 mm |

### 4.2. Patient data

The final example is a set of real patient data. 3D preoperative contrast-enhanced MR angiography (MRA) was performed to delineate endocardial boundaries of the left atrium and pulmonary veins. The voxel size was $0.78125 \times 0.78125 \times 1.5$ mm and 45 slices were used in the experiment. We obtained MRA and 250 EAM data points from the same patient. The EAM data consists of the CARTO points imported from the CARTO-XP, including measurement points as well as ablation points.

After delineating and removing unwanted regions such as the left ventricle (LV) and other small veins, we reconstruct the 3D model as shown in Figure 5 using ITK-SNAP [26] and Matlab software. Afterwards, a two-step registration process is applied for the ICP scheme. First, we perform the landmark registration using three junctions between LA and pulmonary veins: LA-LIPV, LA-LSPV, and LA-RSPV. These points are used for the initial pose of subsequent registration. Second, surface registration using the ICP scheme is performed to refine accuracy furthermore. The resulting image is shown in Figure 6(a).

To validate the proposed algorithm, the optimal surface is reconstructed using 250 EAM data points by incorporating a heart shape prior from preoperative MRA. By minimizing the energy functional, the final result is shown in Figure 6(b), where diamonds (in blue) represent EAM data points, and circles (in red) represent ablation points. Blurred points are located inside. A quantitative evaluation result can be obtained in terms of the mean Euclidean distances of EAM and ablation points from the surface of the left atrium. They are reported in Table 1, which shows that the proposed approach outperforms the ICP method significantly.

## 5. CONCLUSION AND FUTURE WORK

A novel multimodal data integration technique using the level set method for catheter ablation of AF was presented in this paper. This technique enables reconstruction and registration simultaneously using data fitting, regularization, and shape prior energy terms. It provides better performance than the existing ICP method in accuracy. In the proposed framework, the heart-shape model from MRA reconstruction is used as a prior shape knowledge. Thus, we can use the shape information to compensate for insufficient EAM data. Clinically, this technique can improve efficacy and safety of AF ablation by integrating EAM data and 3D imaging data.

Dynamic cardiac shape analysis will make the current integration method more precise and meaningful. We plan to incorporate a richer set of spatiotemporal shape models using dynamic shape information in the future. Besides, we may consider a localized regularization method around the point data to obtain more precise reconstruction.

## REFERENCES

[1] D. P. Zipes and H. J. J. Wellens, "What have we learned about cardiac arrhythmias?" *Circulation*, vol. 102, no. 4, pp. 52–57, 2000.

[2] I. B. Ray and E. Heist, "Treating atrial fibrillation: what is the consensus now?" *Postgraduate Medicine*, vol. 118, no. 4, pp. 47–58, 2005.

[3] C. Pappone, S. Rosanio, G. Oreto, et al., "Circumferential radiofrequency ablation of pulmonary vein ostia: a new anatomic approach for curing atrial fibrillation," *Circulation*, vol. 102, no. 21, pp. 2619–2628, 2000.

[4] J. Dong, H. Calkins, S. B. Solomon, et al., "Integrated electroanatomic mapping with three-dimensional computed tomographic images for real-time guided ablations," *Circulation*, vol. 113, no. 2, pp. 186–194, 2006.

[5] V. Y. Reddy, Z. J. Malchano, G. Holmvang, et al., "Integration of cardiac magnetic resonance imaging with three-dimensional electroanatomic mapping to guide left ventricular catheter manipulation: feasibility in a porcine model of healed myocardial infarction," *Journal of the American College of Cardiology*, vol. 44, no. 11, pp. 2202–2213, 2004.

[6] Z. Malchano, "Image guidance in cardiac electrophysiology," M.S. thesis, Massachusette Institute of Technology, Boston, Mass, USA, June 2006.

[7] P. J. Besl and H. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[8] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proceedings of the 3rd International Conference on 3D Digital Imaging and Modeling (3DIM '01)*, pp. 145–152, Quebec, Canada, May-June 2001.

[9] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The trimmed iterative closest point algorithm," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, vol. 3, pp. 545–548, Quebec, Canada, August 2002.

[10] T. Masuda, K. Sakaue, and N. Yokoya, "Registration and integration of multiple range images for 3-d model construction," in *Proceedings of the 13th International Conference on Pattern Recognition (ICPR '96)*, pp. 879–883, Vienna, Austria, August 1996.

[11] K. Pulli, "Multiview registration for large data sets," in *Proceedings of the 2nd International Conference on 3D Digital Imaging and Modeling (3DIM '99)*, pp. 160–168, Ottawa, Canada, October 1999.

[12] E. K. Heist, J. Chevalier, G. Holmvang, et al., "Factors affecting error in integration of electroanatomic mapping with CT and MR imaging during catheter ablation of atrial fibrillation," *Journal of Interventional Cardiac Electrophysiology*, vol. 17, no. 1, pp. 21–27, 2006.

[13] J. Woo, B.-W Hong, S. Kumar, I. B. Ray, and C.-C J. Kuo, "Joint reconstruction and registration using level sets: application to the computer-aided ablation of atrial fibrillation," in *Proceedings of International Conference Frontiers in the Convergence of Bioscience and Information Technologies*, Jeju, Korea, October 2007.

[14] G. C.-H. Chuang and C.-C. J. Kuo, "Wavelet descriptor of planar curves: theory and applications," *IEEE Transactions on Image Processing*, vol. 5, no. 1, pp. 56–70, 1996.

[15] D. Rogers, *An Introduction to NURBS: With Historic Perspective*, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2001.

[16] N. Amenta and M. Bern, "Surface reconstruction by Voronoi filtering," *Discrete and Computational Geometry*, vol. 22, no. 4, pp. 481–504, 1999.

[17] H. Edelsbrunner, "Shape reconstruction with delaunay complex," in *Proceedings of the 3rd Latin American Symposium on Theoretical Informatics (LATIN '98)*, pp. 119–132, Springer, Campinas, Brazil, April 1998.

[18] A. Dervieux and F. Thomasset, *A Finite Element Method for the Simulation of a Rayleigh-Taylor Instability*, vol. 771 of *Lecture Notes in Mathematics*, Springer, Berlin, Germany, 1979.

[19] A. Dervieux and F. Thomasset, "Multifluid incompressible flows by a finite element method," in *Seventh International Conference on Numerical Methods in Fluid Dynamics*, W. Reynolds and R. MacCormack, Eds., vol. 141 of *Lecture Notes in Physics*, pp. 158–163, Springer, Berlin, Germany, 1980.

[20] S. Osher and J. A. Sethian, "Fronts propagating with curvature dependent speed: algorithms based on the Hamilton-Jacobi formulation," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[21] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points," *Computer Graphics*, vol. 26, no. 2, pp. 71–78, 1992.

[22] H.-K. Zhao, S. Osher, and R. Fedkiw, "Fast surface reconstruction using the level set method," in *Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods, The 8th IEEE International Conference on Computer Vision*, pp. 194–202, Vancouver, BC, Canada, July 2001.

[23] J. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry*, Cambridge University Press, Cambridge, UK, 1998.

[24] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[25] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher, "A variational level set approach to multiphase motion," *Journal of Computational Physics*, vol. 127, no. 1, pp. 179–195, 1996.

[26] P. A. Yushkevich, J. Piven, H. Cody, S. Ho, J. C. Gee, and G. Gerig, "User-guided level set segmentation of anatomical structures with ITK-SNAP," *Insight Jounral*, vol. 1, 2005, special issue on ISC/NA-MIC/MICCAI Workshop on Open-Source Software.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 326464, 8 pages
doi:10.1155/2008/326464

*Research Article*

# Peptide Mimicrying Between SARS Coronavirus Spike Protein and Human Proteins Reacts with SARS Patient Serum

**K.-Y. Hwa,[1, 2, 3, 4] W. M. Lin,[3] Y.-I. Hou,[5] and T.-M. Yeh[5]**

[1] *Department of Molecular Science and Engineering, Center for Biomedical Industries, National Taipei University of Technology, Taipei 106, Taiwan*
[2] *Institute of Polymeric Science, National Taiwan University, Taipei 10617, Taiwan*
[3] *Center for Biomedical Industries, National Taipei University of Technology, Taipei 106, Taiwan*
[4] *Institute of Biomedical Technology, Taipei Medical University, Taipei 110, Taiwan*
[5] *Department of Medical Laboratory Science and Biotechnology, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan*

Correspondence should be addressed to T.-M. Yeh, today@mail.ncku.edu.tw

Received 1 October 2007; Accepted 24 December 2007

Recommended by Daniel Howard

Molecular mimicry, defined as similar structures shared by molecules from dissimilar genes or proteins, is a general strategy used by pathogens to infect host cells. Severe acute respiratory syndrome (SARS) is a new human respiratory infectious disease caused by SARS coronavirus (SARS-CoV). The spike (S) protein of SARS-CoV plays an important role in the virus entry into a cell. In this study, eleven synthetic peptides from the S protein were selected based on its sequence homology with human proteins. Two of the peptides D07 (residues 927–937) and D08 (residues 942–951) were recognized by the sera of SARS patients. Murine hyperimmune sera against these peptides bound to proteins of human lung epithelial cells A549. Another peptide D10 (residues 490–502) stimulated A549 to proliferate and secrete IL-8. The present results suggest that the selected S protein regions, which share sequence homology with human proteins, may play important roles in SARS-CoV infection.

## 1. INTRODUCTION

Severe acute respiratory syndrome (SARS) is a new emerging infectious disease, which was first reported in China in 2002 and was rapidly spreading all over the world in 2003 [1, 2]. The disease was transmitted by droplets and close contact. Patients develop persistent fever, dry cough, progressive radiographic changes of chest, and lymphopenia once infected. Despite treatment, about 10–15% of the patients would die due to the acute respiratory distress [3–6]. A novel coronavirus (SARS-CoV) was isolated from SARS patients [7–9]. SARS-CoV is a positive-stranded RNA virus with an envelop. The genome of SARS-CoV is around 29,727 nucleotides in length. The sequence was annotatedin silico [10]. Comparative genomic studies using the in silico annotated proteins have suggested that SARS virus belongs to a new group of coronavirus.

According to the genomic sequence of SARS-CoV, it is predicted that there are several structural proteins can be produced by SARS-CoV including spike (S), envelop (E), membrane (M), and nucleocapsid. Spike protein is very important in the binding and fusion of coronavirus to the host cells [11, 12]. The S protein of SARS-CoV has 1255 amino acids in length and 23 potential N-linked glycosylation sites. The amino terminus of the SARS-CoV S protein contains a short type 1 signal sequence composed of hydrophobic amino acids that are presumably removed during cotranslational transport through the endoplasmic reticulum. The carboxyl terminus consists of a transmembrane domain and a cytoplasmic tail rich in cysteine residues. The majority of protein (residues 12–1195) is outside the virus particle, which can be divided into amino-terminal S1 and carboxyl-terminal S2 domain. The S1 domain (residues 12–672) binds to the host cell receptor, angiotensin-converting enzyme 2 (ACE2), while the S2 domain is responsible for membrane fusion [13–15]. Monoclonal antibodies against S1 domain can block the receptor binding and contain potent neutralization activity against SARS-CoV [16]. However, peptides derived from S2 domain can also inhibit SARS-CoV infection [12].

Molecular mimicry, which is defined as similar structures shared by molecules from dissimilar genes or by their protein products, is a general strategy for pathogens to infect host cells and has been proposed as a pathogenic mechanism for autoimmune disease [17]. Therefore, identification of the molecular mimic regions of pathogen may be helpful to understand the disease induced by that pathogen. At present, it is unclear whether molecular mimicry occurs between SARS-CoV S proteins and human peptides. We have approached this question using computer to analyze the sequence of spike protein of SARS-CoV and select regions that share the sequence homology with human proteins. The criteria for the selection of potential regions include antigenic analysis and surface accessibility. In this study, we find that several regions of the S protein share sequence homology with human proteins. Synthetic peptides, which represent some of these regions, were synthesized to understand their roles in SARS-CoV infection.

## 2. MATERIALS AND METHODS

### 2.1. Peptide prediction and synthesis

Publically available human and coronavirus genome sequences at the National Center for Biotechnology Information (Md, USA) were used for in silicoprediction. Algorithms predicting immunogenicity, second structure prediction, protein topology analysis, and hydrophobicity were conducted to design the tested peptides. Immunogenic viral peptides were calculated based on the algorithm developed by Kolaskar and Tongaonkar [18]. The algorithm is based on a table constructed from the occurrence of amino acid residues in experimentally known antigenic epitopes. The reported accuracy of the method is about 75% [18]. In silico secondary structural analyses of spike protein were performed based on PHD [19] and PREDATOR [20] algorithms. Protein topology prediction was based on the algorithm developed by TMHMM [21]. Hydrophobicity of the peptides was calculated based on the algorithm HMOMENT [22]. Similarity searches between S protein and human genome database were performed by using BLASTP [23] with BLOSUM 62. Extra amino acid residues were added at either N- or C-terminus to keep the hydrophobic amino acid content below 50%. Peptides with high hydrophobicity are difficult to be tested in biochemical experiments since most of in vitro assays are conducted in aqueous buffers. Also, on average, one charged residue is added for every five amino acids. Multiple antigen peptides were synthesized by CytoMol Corp (Mountain View, Calif, USA). In addition, bradykinin and angiotensin I (Ang I) were purchased from Sigma (St. Louis, Mo, USA).

### 2.2. SARS patient sera

SARS patient sera were collected by the Center for Disease Control, Department of Health (Taipie, Taiwan) from March to June, 2003. Diagnosis of SARS was based on the clinicalcriteria established by the World Health Organization (WHO). Patients with SARS-CoV were confirmed by laboratory methods, including viral antigen detection, RT-PCR, and serologic methods. Ten SARS patient sera collected at the convalescent stage (≥20 days after disease onset) were included in this study. Ten normal sera from healthy individuals were used as controls.

### 2.3. Enzyme-linked immunosorbent assay (ELISA)

Antibodies against peptides in human sera were detected by solid-phase capture technique using individual peptide-coated plates. ELISA plate was coated with or without 50 $\mu$L peptides (100 $\mu$g/mL) per well and blocked by 1% bovine serum albumin (BSA) in 0.05% Tween-20 in phosphate-buffered saline (PBS) for 1 hour at room temperature. Test serum samples were 1 : 100 diluted and added to the plate for 2 hours at room temperature. After incubation, the ELISA plate was washed with 0.05% Tween-20 in PBS for three times. The bound antibodies were detected by horseradish peroxidase- (HRP-) conjugated antihuman immunoglobulin antibodies (Sigma Aldrich, St. Louis, Mo, USA) and peroxidase substrate, TMB (Promega, Madison, Wiss, USA). The absorbance was measured using the Vmax microplate reader (Molecular Devices Corporation, Sunnyvale, Calif, USA) at 450 nm. Antibodies against peptides in mouse sera were assayed by ELISA as in human sera except HRP-conjugated antimouse immunoglobulin antibodies (Sigma Aldrich) was used to detect bound antibodies.

### 2.4. Cell culture

Human lung adenocarcinoma cell line A549 and Vero cells were grown in DMEM supplemented with 10% heat-inactivated FCS, 2 mM L-glutamate, and 50 ng/mL gentamycin. Cells were incubated in $CO_2$ incubator at 37°C with 5% $CO_2$ in a humidified atmosphere. For immunofluorescent microscopy observation, monolayers of A549 cells were cultured on sterile glass slides before the experiment.

### 2.5. Mice immunization

Six- to eight-week-old female BALB/c mice were used in this study. These mice were originally purchased from Jackson Laboratory (Bar Harbor, Me, USA) and bred in the Laboratory Animal Center, National Cheng Kung University (Tainan, Taiwan). Synthetic peptides (1 mg/mL) were emulsified with complete Freund's adjuvant and injected intraperitoneally into BALB/c mice. Mice were boosted with the same peptide in PBS (50 $\mu$g/mouse) intraperitoneally two weeks after priming. Sera were collected from the axially plexus of the mice at different time intervals and tested for the presence of antibody against peptides by ELISA as mentioned above. Significant increase of antibody titer (greater than 4 folds) against immunized-peptide was found in mouse hyperimmune sera as compared to normal sera after boosting.

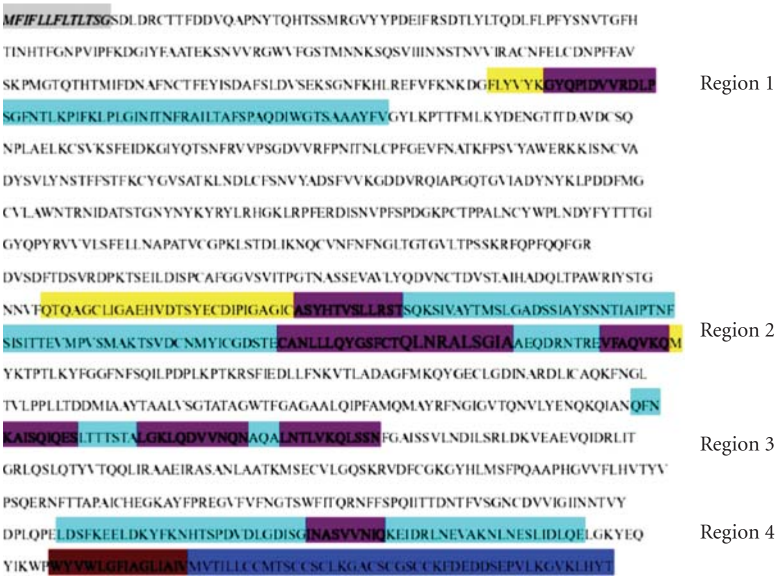

MFIFLLFLTLTSGSDLDRCTTFDDVQAPNYTQHTSSMRGVYYPDEIFRSDTLYLTQDLFLPFYSNVTGFH
TINHTFGNPVIPFKDGIYFAATEKSNVVRGWVFGSTMNNKSQSVIIINNSTNVVIRACNFELCDNPFFAV
SKPMGTQTHTMIFDNAFNCTFEYISDAFSLDVSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLP
SGFNTLKPIFKLPLGINITNFRAILTAFSPAQDIWGTSAAAYFVGYLKPTTFMLKYDENGTITDAVDCSQ
NPLAELKCSVKSFEIDKGIYQTSNFRVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVA
DYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMG
CVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTPPALNCYWPLNDYFYTTTGI
GYQPYRVVVLSFELLNAPATVCGPKLSTDLIKNQCVNFNFNGLTGTGVLTPSSKRFQPFQQFGR
DVSDFTDSVRDPKTSEILDISPCAFGGVSVITPGTNASSEVAVLYQDVNCTDVSTAIHADQLTPAWRIYSTG
NNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLLRSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNF
SISITTEVMPVSMAKTSVDCNMYICGDSTECANLLLQYGSFCTQLNRALSGIAAEQDRNTREVFAQVKQM
YKTPTLKYFGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGECLGDINARDLICAQKFNGL
TVLPPLLTDDMIAAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFN
KAISQIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLIT
GRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQAAPHGVVFLHVTYV
PSQERNFTTAPAICHEGKAYFPREGVFVFNGTSWFITQRNFFSPQIITTDNTFVSGNCDVVIGIINNTVY
DPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQ
YIKWPWYVWLGFIAGLIAIVMVTILLCCMTSCCSCLKGACSCGSCCKFDEDDSEPVLKGVKLHYT
Region 1
Region 2
Region 3
Region 4

Figure 1: Sequence analysis of S protein. Putative S protein amino acid sequence was analyzed to find immunogenic regions (yellow regions) and pathogenic regions (regions shared sequence homology with human proteins, blue regions). Purple regions are both immunogenic and pathogenic regions. The grey region is the leader sequence and the brown region is the transmembrane region.

```
  S protein  893 VTQNVLYENQKQIAN----------QFNKAISQIQESLTTTSTALGKLQDVV-NQNAQAL 941
                 +T  V   N  Q AN          Q NKAI  ++Q +++TT   L  LQD + NQ A A+
 Angrgm-52   126 ITAGVALHNSIQNANAILTLKDSIRQSNKAIQELQTAMSTTVVVLNALQDQINNQLAPAI  185

  S protein  942 NTLVKQLSSN 951
                 N+L  Q+ +N
 Angrgm-52   186 NSLGCQVVAN 195
```

(a)

```
 S protein    GYQPYRVVVLS
              | : : | : | . . . : .
Bradykinin    GFSPFRSSRIG
```

(b)

Figure 2: Sequence homology between S protein, (a) human Angrgm-52, and (b) bradykinin. Spike protein sequence (protein databank NP_828851) was compared with sequences from all nonredundant GenBank CDS translations, PDB, SwissProt, PIR, and PRF of human, by using blastp (NCBI, NIH, USA). For Panel (a), BLASTP was used to calculate the similarity between S protein and human Angrgm-52, with matrix set at default. For an identical residue, one letter symbol of amino is shown between the two sequences, and for a conservative substation, "+" is shown. For Panel (b), pairwise alignment was calculated based on Smith-Water local alignment with matrix set at BLOSUM 45. "|" is annotated for identical residues; ":" and "." are for similar residues.

### 2.6. Immunofluorescent stain

Mouse hyperimmune sera against peptide D08 were incubated with A549 cells at 4°C for 1 hour. After washing three times with PBS, cells were incubated with 1 mL of 1 $\mu$g/mL FITC-conjugated antimouse IgG (Jackson ImmunoResearch Laboratories Inc., West Grove, Pa, USA) at 4°C for 1 hour and washed again with PBS. The immunofluorescent stain of cells was observed by fluorescent microscopy.

### 2.7. SDS-PAGE and western blot analysis

Proteins in the cell lysate of A549 were separated by 12% SDS-PAGE and transferred to nitrocellulose sheets as described previously [24]. Proteins recognized by normal or peptide D08 hyperimmune mice sera were detected using HRP-conjugated antimouse immunoglobulin antibodies (Sigma Aldrich) and substrate.

### 2.8. Cell proliferation

Vero E6 ($4 \times 10^4$) and A549 cells ($5 \times 10^3$) were incubated with different doses of synthetic peptides as indicated for 72 hours. Cell proliferation was detected using commercial XTT assay (Roche Diagnostics, Indianapolis, Ind, USA).

### 2.9. IL-8 assay

The IL-8 production was assessed by commercial ELISA kits (R&D systems, Minneapolis, Minn, USA) according to the manufacturer's instructions. Briefly, A549 cells ($1 \times 10^5$) were cultured alone or with different doses of peptides for 48 hours. Culture supernatants were collected after incubation, added to precoated ELISA plates, and incubated for 2 hours at 37°C. Plates were washed four times with the washing buffer. The bound IL-8 was detected by HRP-conjugated antibodies and substrate. The developed color was read by the Vmax microplate reader (Molecular Devices, Calif, USA). The concentration of IL-8 was calculated according to the standard curve.

### 2.10. Statistical analyses

Data are expressed as mean ± standard deviation (SD). The levels of significance for the differences between groups were

Table 1: Amino acid sequence of the eleven synthetic peptides *Extra amino acid residues which were indicated by italic letters were added at either N- or C-terminus to decrease the hydrophobicity.

| Peptide | Amino acid positions | Amino acid sequence | No. of amino acids |
|---|---|---|---|
| D01 | 199–210 | GYQPIDVVRDL*G* | 12 |
| D02 | 658–669 | ASYHTVSLLRST*SQK* | 15 |
| D03 | 733–744 | *EEG*NLLLQYGSFCTQ | 15 |
| D04 | 745–753 | *EE*LNRALSGIAAGQ | 13 |
| D05 | 763–770 | VFAQVKQM | 8 |
| D06 | 911–919 | KAISQIQES*LTTE* | 13 |
| D07 | 927–937 | *G*LGKLQDVVNQN*GE* | 14 |
| D08 | 942–951 | *A*LNTLVKQLSSN | 12 |
| D09 | 1154–1162 | INASVVNIQ*K* | 10 |
| D10 | 490–502 | GYQPYRVVVLSFEE | 14 |
| D11 | 306–317 | *G*FRVVPSGDVVR*F* | 13 |

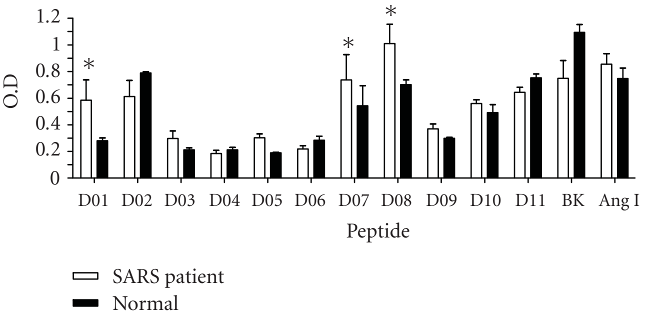
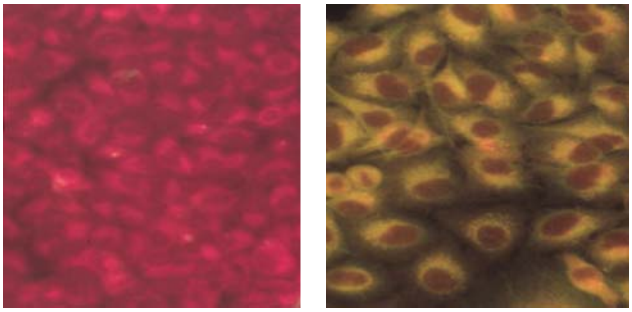


Figure 3: Antibody binding activity of SARS patients' sera to different peptides. Sera of SARS patients at convalescent stage as well as normal controls were collected as described in Section 2. Antibodies binding to different peptides were detected by ELISA as described in Section 2. BK represents bradykinin. "∗" indicates $P < .05$.



(a) (b)

Figure 4: Immunofluorescent staining of mouse hyperimmune sera against A549 cells. A549 cells were grown on the slides and stained with secondary antibody alone (a) or mouse hyperimmune sera against D08 peptide (b) as described in Section 2.

analyzed using Student's $t$-test. A value of $P < .05$ was considered to be significant.

## 3. RESULTS

### 3.1. Search for molecular mimic regions in S protein

The whole amino acid sequence of spike protein was analyzed to find out the potential immunogenic regions and the regions shared sequence homology with human proteins, which is defined as the pathogenic regions. As shown in Figure 1, there are 4 pathogenic regions. Region 1 (residues 199–254), region 2 (residues 658–715), region 3 (residues 893–951), and region 4 (residues 1127–1184) have shared sequence homology with hydroxyacid oxidase, human golgi autoantigen, Angrgm-52, and pallidin, respectively. Among these regions, region 3 with homology to human Angrgm-52 has the highest score (with 34% identities and 48% similarity of conservative substitutions). Its sequence comparison with angrgm-52 is shown in Figure 2(a). In addition, because des-Arg bradykinin and Ang I are the substrates for ACE2 [25], we also compared the sequence of S protein against bradykinin (RPP*GFSPFR*) and Ang I (DRVYIHPFHL) and found that residues 490–502 (*GYQPYR*VVVLSFEE) of S protein showed sequence homology with bradykinin as indicated by bold letters here and in Figure 2(b). The identity score is 27%; and the similarity score from conservative substitutions is 36%.

### 3.2. Screen for peptides recognized by SARS patients' sera

Eleven peptides (D01–D11, see Table 1), which represent those pathogenic regions were synthesized as well as bradykinin and Ang I were tested to see whether those peptides can be recognized by SARS patients' sera. The peptides were designed based on the algorithms predicting immunogenicity, second structure, protein topology, and hydrophobicity. Our goal is to select for peptides with high immunogenicity, with location on the protein surface, and with low hydrophobicity. The designed peptide sequences were synthesized and tested with clinical samples of SARS patient sera. A significant increase of SARS patients' sera binding to peptide D01, D07, and D08 was found as compared to the binding of normal sera (see Figure 3).

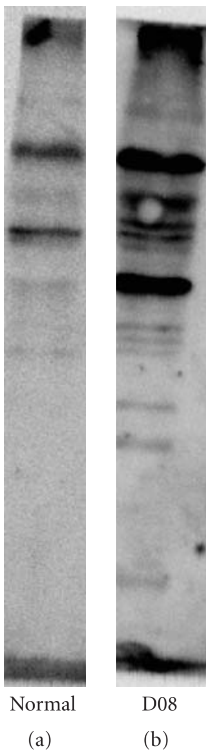


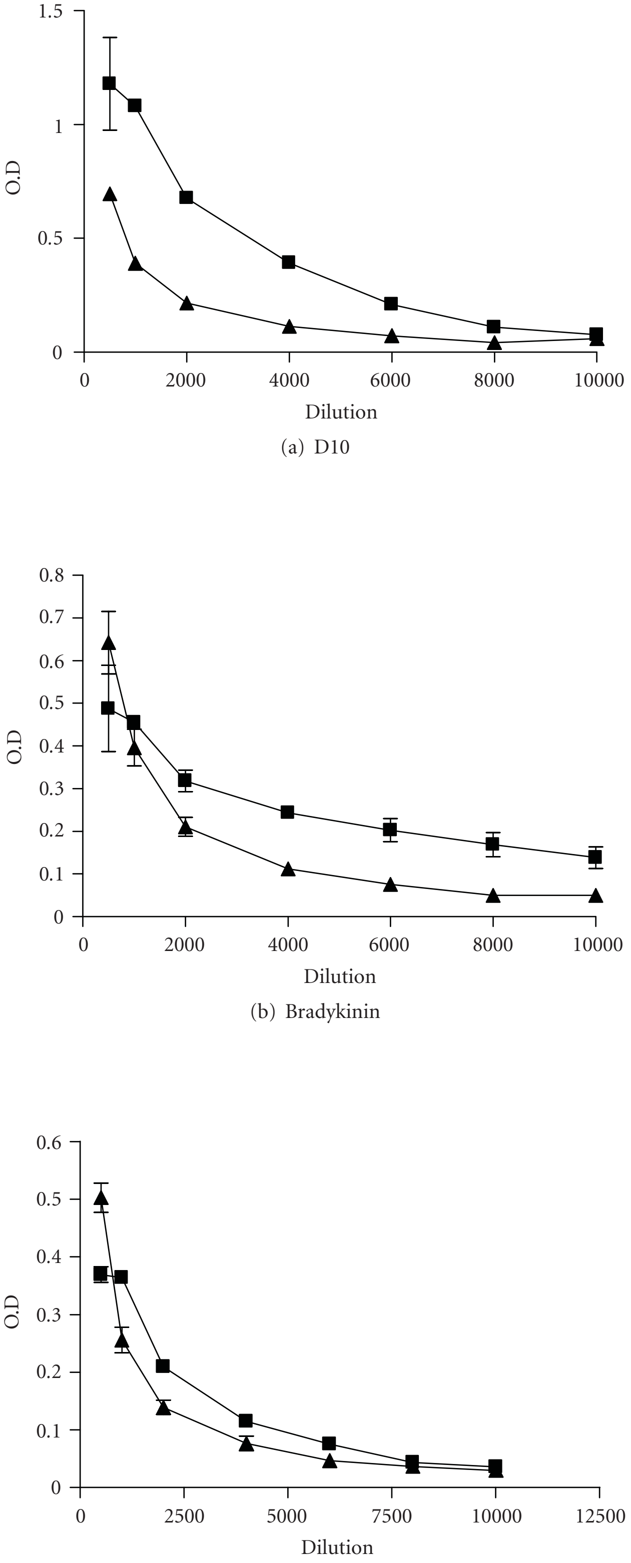
Figure 5: Mouse hyperimmune sera against D08 peptide recognize proteins in the A549 cell lysate. Proteins in the cell lysate of A549 cells were separated by SDS-PAGE and transferred to membrane as described in Section 2. Western blots against this membrane using normal mouse sera (a) or hyperimmune sera against D08 (b) are shown.

### 3.3. *Hyperimmune sera against D08 crossreacted with A549 cells*

To test whether synthetic peptides D01, D07, and D08 can induce antibodies crossreacted with human proteins, we immunized mice with these peptides to generate hyperimmune sera against these peptides. We found hyperimmune sera against D08 can bind to the cytoplasmic region of human A549 cells as demonstrated by immunofluorescent stain (see Figure 4). Using Western blot analysis, hyperimmune sera against D08 could recognize more bands in A549 cell lysate as compared to normal mice sera (see Figure 5). In addition, hyperimmune sera against D07, but not D01, showed similar crossreactivity to A549 cells as hyperimmune sera against D08 did (data not shown).

### 3.4. *Hyperimmune sera against D10 crossreacted with bradykinin*

To test whether synthetic peptides D10, indeed, can induce antibodies crossreactive with bradykinin and Ang I, we immunized mice with D10 peptides to generate hyperimmune sera against this peptide. Significant increase of antibodies against D10 was found in D10 hyperimmune sera, which could crossreact with bradykinin, but not with Ang I-coated plates (see Figure 6).

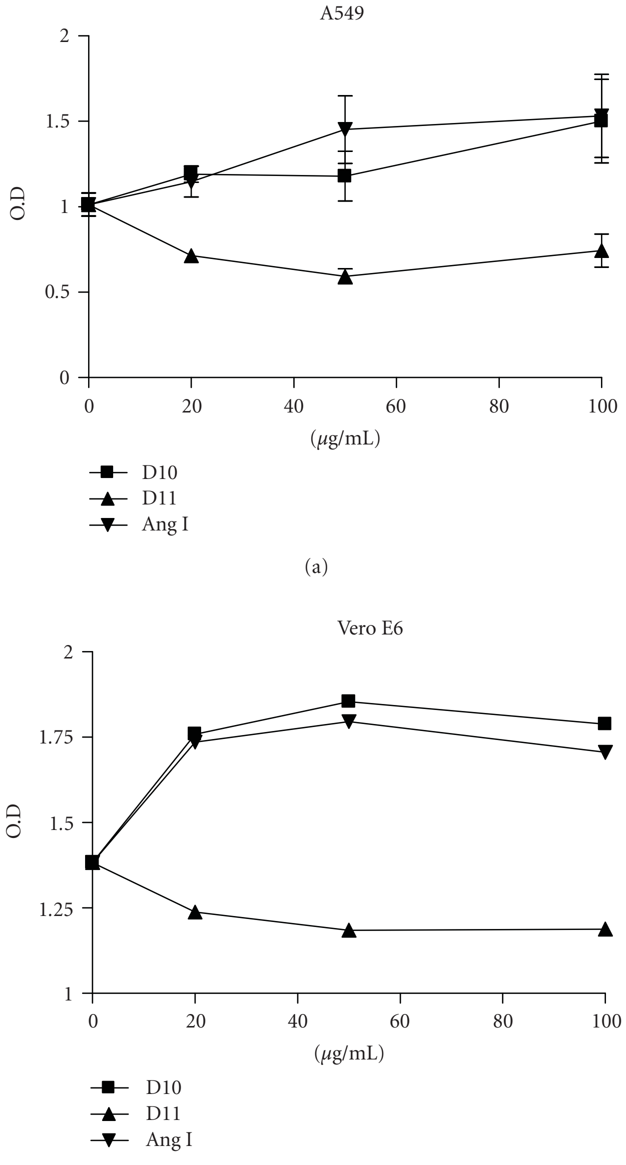


(a) D10

(b) Bradykinin

(c) Angiotensin I

Figure 6: The crossreactivity of D10 antibody with bradykinin and Ang I. Hyperimmune sera from D10 immunized mice (■) or normal mice sera (▲) were diluted as indicated and reacted with D10-, bradykinin-, or Ang I-coated ELISA plates as indicated. Bound antibodies were detected as described in Section 2. Data represents the mean ± SD of triplicates.



(a)

(b)

Figure 7: Cell proliferation induced by D10 peptide and Ang I. Vero and A549 cells were incubated with different doses of peptides as indicated. Cell proliferation was detected after 72 hours of incubation by XTT assay as described in Section 2. Data represents the mean ± SD of triplicates.

### 3.5. *Peptide D10 induced IL-8 secretion and cell proliferation of A549 cells*

To understand whether D10 has similar biological activity as Ang I, we incubated Vero cells and lung epithelial A549 cells with D10, Ang 1, or control peptide D11. Both Vero and A549 cells were induced to proliferate in the presence of D10 and Ang I but not the control peptide (see Figure 7). In addition, D10 and Ang I could also induce chemokine IL-8 production of A549 cells (see Figure 8).

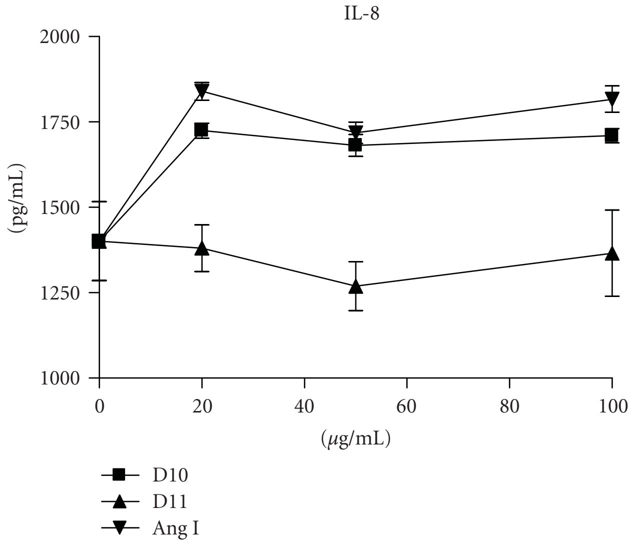


Figure 8: IL-8 production of A549 cells induced by D10 peptide and Ang I. A549 cells ($1 \times 10^5$) were incubated with or without peptides for 48 hours. The levels of IL-8 in the culture supernatants were assayed as described in Section 2. Data represents the mean ± SD of triplicates.

## 4. DISCUSSION

In this study, we have identified four pathogenic regions of SARS-CoV S protein which share sequence homology with different human proteins. Among them, pathogenic region 3 (residues 893–941), which shares sequence homology with Angrgm-52 (GenBank accession no. AAL62340), a novel gene upregulated in human mesangial cells stimulated by angiotensin II, may deserve further investigation. Peptides D07 and D08 of this region were recognized by the sera of SARS patient indicating that this region is immunogenic and can be recognized by the immune system during SARS-CoV infection. Murine hyperimmune sera against peptides D07 or D08 were able to bind to recombinant S2 but not S1 domain of S protein (data not shown). In addition, hyperimmune sera against D07 or D08 also bounded to the cytoplasmic region of A549 cells and recognized several proteins in the A549 cell lysate. These results indicate that regions represented by D07 and D08 are immunogenic and may induce autoantibodies. However, further study is required to understand the biological function of these regions and the role of their antibodies in the pathogenesis of SARS-CoV infection.

In addition to D07 and D08 peptides, we also noticed that D10 peptide which represents residues 490–502 of S1 domain contained some interesting activities. The D10 peptide, which shared sequence homology with bradykinin, was able to generate antibodies crossreactive with bradykinin. In addition, D10 peptide could stimulate A549 to produce IL-8 and proliferation as Ang I did. These results suggest that the region of D10 in S protein may bind to Ang I receptor, ACE2, and may be involved in the binding of SARS-CoV to ACE2. This is consistent with the previous report, which indicates that residues 318–510 of S1 domain can bind to ACE2 [25] and is similar to the receptor binding domain of the HCoV-229E, which is within a fragment containing residues 407 to 547 [26]. Therefore, region 490–502 of S1 domain may be involved in the receptor binding domain of SARS-CoV.

In summary, our results suggest that molecular mimicry occurs between SARS-CoV and host proteins. Motifs shared sequence homology with host proteins of SARS-COV may be involved in the binding and fusion of SARS-CoV to host cells. Antibody against these motifs may contain neutralization activity against SARS-CoV infection or participate in the immunopathogenesis induced by SARS-CoV. As reported previously, SARS-CoV, like influenza, can inhibit the host's corticosteroid stress response via a molecular mimicry strategy [27]. Our studies on the mimicry motifs of S protein, which is involved in the virus, entry may provide alternative approaches to disrupt the infection of SARS-CoV, similar to the previous studies on the virus entry [28, 29].

## ACKNOWLEDGMENT

This work was supported by Grant no.NSC92-2751-B006-Y from the National Science Council, Taipei, Taiwan.

## REFERENCES

[1] M. D. Christian, S. M. Poutanen, M. R. Loutfy, M. P. Muller, and D. E. Low, "Severe acute respiratory syndrome," *Clinical Infectious Diseases*, vol. 38, pp. 1420–1427, 2004.

[2] T. Kuiken, R. A. M. Fouchier, M. Schutten, et al., "Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome," *Lancet*, vol. 362, no. 9380, pp. 263–270, 2003.

[3] N. Lee, D. Hui, A. Wu, et al., "A major outbreak of severe acute respiratory syndrome in Hong Kong," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1986–1994, 2003.

[4] S. M. Poutanen, D. E. Low, B. Henry, et al., "Identification of severe acute respiratory syndrome in Canada," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1995–2005, 2003.

[5] K. W. Tsang, P. L. Ho, G. C. Ooi, et al., "A cluster of cases of severe acute respiratory syndrome in Hong Kong," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1977–1985, 2003.

[6] R. P. Wenzel and M. B. Edmond, "Managing SARS amidst uncertainty," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1947–1948, 2003.

[7] C. Drosten, S. Günther, W. Preiser, et al., "Identification of a novel coronavirus in patients with severe acute respiratory syndrome," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1967–1976, 2003.

[8] T. G. Ksiazek, D. Erdman, C. S. Goldsmith, et al., "A novel coronavirus associated with severe acute respiratory syndrome," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1953–1966, 2003.

[9] P. A. Rota, M. S. Oberste, S. S. Monroe, et al., "Characterization of a novel coronavirus associated with severe acute respiratory syndrome," *Science*, vol. 300, no. 5624, pp. 1394–1399, 2003.

[10] M. A. Marra, S. J. M. Jones, C. R. Astell, et al., "The genome sequence of the SARS-associated coronavirus," *Science*, vol. 300, no. 5624, pp. 1399–1404, 2003.

[11] A. Bonavia, B. D. Zelus, D. E. Wentworth, P. J. Talbot, and K. V. Holmes, "Identification of a receptor-binding domain of the spike glycoprotein of human coronavirus HCoV-229E," *Journal of Virology*, vol. 77, no. 4, pp. 2530–2538, 2003.

[12] B. J. Bosch, B. E. Martina, R. van der Zee, et al., "Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 22, pp. 8455–8460, 2004.

[13] Y. He, Y. Zhou, H. Wu, et al., "Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines," *Journal of Immunology*, vol. 173, no. 6, pp. 4050–4057, 2004.

[14] W. Li, M. J. Moore, N. Vasllieva, et al., "Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus," *Nature*, vol. 426, no. 6965, pp. 450–454, 2003.

[15] S. K. Wong, W. Li, M. J. Moore, H. Choe, and M. Farzan, "A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2," *Journal of Biological Chemistry*, vol. 279, no. 5, pp. 3197–3201, 2004.

[16] J. Sui, W. Li, A. Murakami, et al., "Potent neutralization of severe acute respiratory syndrome (SARS) coronavirus by a human mAb to S1 protein that blocks receptor association," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 8, pp. 2536–2541, 2004.

[17] M. B. A. Oldstone, "Molecular mimicry and immune-mediated diseases," *The FASEB Journal*, vol. 12, no. 13, pp. 1255–1265, 1998.

[18] A. S. Kolaskar and P. C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," *FEBS Letters*, vol. 276, no. 1-2, pp. 172–174, 1990.

[19] B. Rost, G. Yachdav, and J. Liu, "The predict protein server," *Nucleic Acids Research*, vol. 32, Web Server Issue, pp. 321–326, 2004.

[20] D. Frishman and P. Argos, "Incorporation of long-distance interactions into a secondary structure prediction algorithm," *Protein Engineering*, vol. 9, no. 2, pp. 133–142, 1996.

[21] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[22] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, "The hydrophobic moment detects periodicity in protein hydrophobicity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 1, pp. 140–144, 1984.

[23] W. Gish and D. J. States, "Identification of protein coding regions by database similarity search," *Nature Genetics*, vol. 3, no. 3, pp. 266–272, 1993.

[24] T. M. Yeh, H. C. Chang, C. C. Liang, J. J. Wu, and M. F. Liu, "Deoxyribonuclease-inhibtory antibodies in systemic lupus erythematosus," *Journal of Biomedical Science*, vol. 10, no. 5, pp. 544–551, 2003.

[25] S. K. Wong, W. Li, M. J. Moore, H. Choe, and M. Farzan, "A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2," *Journal of Biological Chemistry*, vol. 279, no. 5, pp. 3197–3201, 2004.

[26] J. J. Breslin, I. Mork, M. K. Smith, et al., "Human coronavirus 229E: receptor binding domain and neutralization by soluble receptor at 37°C," *Journal of Virology*, vol. 77, no. 7, pp. 4435–4438, 2003.

[27] R. Wheatland, "Molecular mimicry of ACTH in SARS—implications for corticosteroid treatment and prophylaxis," *Medical Hypotheses*, vol. 63, no. 5, pp. 855–862, 2004.

[28] R. Y. Kao, W. H. Tsui, T. S. Lee, et al., "Identification of novel small-molecule inhibitors of severe acute respiratory syndrome-associated coronavirus by chemical genetic," *Chemistry & Biology*, vol. 11, no. 9, pp. 1293–1299, 2004.

[29] D. P. Han, A. Penn-Nicholson, and M. W. Cho, "Identification of critical determinants on ACE2 for SARS-CoV entry and development of a potent entry inhibitor," *Virology*, vol. 350, no. 1, pp. 15–25, 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 856314, 8 pages
doi:10.1155/2008/856314

*Research Article*

# Focal Point Theory Models for Dissecting Dynamic Duality Problems of Microbial Infections

**S.-H. Huang,[1] W. Zhou,[2] and A. Jong[1]**

[1] *Childrens Hospital Los Angeles, University of Southern California, Los Angeles, CA 90027, USA*
[2] *HRL Laboratories, LLC., Malibu, CA 90265, USA*

Correspondence should be addressed to S.-H. Huang, shhuang@hsc.usc.edu

Received 31 August 2007; Accepted 19 December 2007

Recommended by Daniel Howard

Extending along the dynamic continuum from conflict to cooperation, microbial infections always involve symbiosis (Sym) and pathogenesis (Pat). There exists a dynamic Sym-Pat duality (DSPD) in microbial infection that is the most fundamental problem in infectomics. DSPD is encoded by the genomes of both the microbes and their hosts. Three focal point (FP) theory-based game models (pure cooperative, dilemma, and pure conflict) are proposed for resolving those problems. Our health is associated with the dynamic interactions of three microbial communities (nonpathogenic microbiota (NP) (Cooperation), conditional pathogens (CP) (Dilemma), and unconditional pathogens (UP) (Conflict)) with the hosts at different health statuses. Sym and Pat can be quantitated by measuring symbiotic index (SI), which is quantitative fitness for the symbiotic partnership, and pathogenic index (PI), which is quantitative damage to the symbiotic partnership, respectively. Symbiotic point (SP), which bears analogy to FP, is a function of SI and PI. SP-converting and specific pathogen-targeting strategies can be used for the rational control of microbial infections.

## 1. INTRODUCTION

Infectious diseases caused by bacterial, viral, fungal, or parasitic pathogens continue to be the leading cause of morbidity and mortality worldwide despite the availability of effective antimicrobial agents and vaccines over the last fifty years or more [1]. The continual emergence of previously undescribed new pathogens, reemergence of old pathogens, and the rising crisis of antibiotics resistance will certainly heighten the global impact of microbial infections in the 21st century. These problems are mainly due to inadequate knowledge of the dynamic duality relationships between symbiosis (Sym) and pathogenesis (Pat) in microbial infections [2]. The term symbiosis, which may have many variations on its definition, in this paper refers to living together through a close and prolonged association between two or more organisms of different species [3, 4]. Duality is defined as different ways of looking at the same thing [5]. There are two major limitations inherent in the conventional theories of microbial infection. On the one hand, in the past century, biology and medicine including infectious diseases have been dominated by the reductionistic approaches. Focusing research on individual virulence genes and the important pathogens has been the traditional approach to human infectious diseases. On the other hand, as Joshua Lederberg pointed out [6, 7], medical science is imbued with the Manichaean view of the microbe-human host relationship: “we good; they evil.” Almost all broad-spectrum antimicrobial agents, which are in the best interest of pharmaceutical industries, kill both the good microbes as well as the bad germs. Even though narrow-spectrum antiinfective agents are not “narrow” for pathogens, they also target both the good and bad microorganisms with a limited range of species.

Animals and plants are continually infected by an extensive diversity of symbiotic or invading organisms including bacteria, virus, fungus, or parasites. Infection of bacteria by phages started long before the emergence of animals and plants [8]. Microbial infection is an evolutionary paradigm which is associated with coevolution between hosts and microbes [6, 7, 9]. This coevolution can be defined as the process of reciprocal and dynamic genetic changes in two or more species [2]. The conventional wisdom in medicine holds that microbial infection is a pathogenic process in

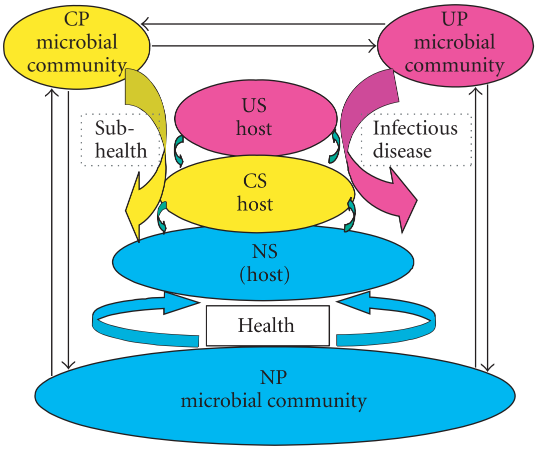


Figure 1: Schematic representation of interactions of three microbial communities (nonpathogenic (NP), conditional pathogenic (CP), and unconditional pathogenic (UP)) with the hosts at three different health statuses (nonsusceptibility (NS), conditional susceptibility (CS), and unconditional susceptibility (US)).

which a pathogen enters, establishes itself, and multiplies in the host [10]. The emphasis is on the antagonism or conflict, not the mutualism. This represents "zero-sum thinking"—the belief that if one player gains, other player must inevitably lose. Methods and concepts of the zero-sum game theory have proved successful in studying the strategy of pure conflict. The most challenging issue in infectious diseases is how to dissect the dynamic Sym-Pat duality (DSPD) in microbial infections using infectomics and mathematics such as focal point- (FP-) based game theory. Game theory, defined in the broadest sense, is the study of the strategies of conflict, cooperation, and mixed situations in which both coexist. Generally, there are multiple equilibria in a game. Thomas Schelling's concept of focal point, which is an equilibrium usually standing out from the others, addressed the crucial question of how to interpret the multiplicity of equilibria [11, 12]. Focal point, the principal component of Schelling's game theory, is a convergence point of expectations about actions in a game. This article attempts to enlarge the scope and application of focal point game theory in microbial infections, extending from the zero-sum games to the nonzero-sum games.

## 2. DEFINITIONS AND METHODS

### 2.1. *Three-community principle of microbial infections*

Our health is associated with the dynamic interactions of three microbial communities [2] (nonpathogenic microbiota (NP), conditional pathogens (CP), and unconditional pathogens (UP)) with the hosts at three different health statuses (nonsusceptibility (NS), conditional susceptibility (CS), and unconditional susceptibility (US)) (see Figure 1). NP is the major microbial community which forms a healthy symbiotic "superorganism" with the hosts. The ecology and evolution of NP-NS interaction are essential and fundamental for health. From birth to death, we share a benign coexistence with a vast, complex, and dynamic consortium of microbes. Most of our microbial commensals reside in our gastrointestinal (GI) track packed with up to 100 trillion ($10^{14}$) microbes [1, 13]. The GI tract harbors a rich microbiota of >600 different bacterial species. Some of these microorganisms have important health functions. These include stimulating the immune system, protecting the host from microbial invasion, and aiding digestion. The gut microbiota, which is essential for human homeostasis, is established rapidly after birth and remains relatively stable throughout the life [1]. The GI mucosa provides a protective interface between the internal environment and the constant external challenge from food-derived antigens and microbes. CP and UP are minor microbial communities that mainly contribute to the pathogenesis of microbial diseases. The distinction between the commensal and the pathogen in the CP community can be blurred because they may cause diseases under certain sub-health conditions of the hosts, or in immunocompromised hosts. For example, pneumococcus, meningococcus, and Haemophilus bacteria regularly exist as part of the normal microbiota of the host respiratory track and are mostly carried asymptomatically despite the fact that they can cause well-defined diseases [14, 15]. Microbes in the CP community dynamically evolve in two opposite directions, which are toward either the NP (more cooperative or mutualistic) or UP (more pathogenic) microbial community. Microbes with high pathogenicity belong to the UP microbial community. The three microbial communities and three statuses of the hosts are subjected to dynamic reciprocal changes driven by transfer of genetic materials.

### 2.2. *Dynamic duality relationships between Sym and Pat in microbial infections*

Extending along the dynamic continuum from conflict to cooperation, microbial infections always involve symbiosis and pathogenesis, which are two fundamental components of the host-microbe interactions (see Figure 2). There exists a dynamic Sym-Pat duality in microbial infection, which is the most fundamental issue of infectomics [2]. DSPD is reflected in the genotypic and phenotypic infectomes, which are encoded by the genomes of both the microbes and their hosts [2]. The opposition and unity of Sym and Pat are indispensable, and the academic viewpoint that the unity of opposites of Sym and Pat gives impetus to the development of microbial infection is considered as the core idea and radical principle of the duality representations of microbial infections. In certain circumstances and at a certain stage of the development of microbial infection, each of the two aspects of Sym and Pat will transform from antagonism into mutualism or from mutualism into antagonism. Sym and Pat can be quantitated by measuring symbiotic index (SI), which is quantitative fitness for the symbiotic partnership, and pathogenic index (PI), which is quantitative damage to the symbiotic partnership, respectively. The most crucial studies are to identify

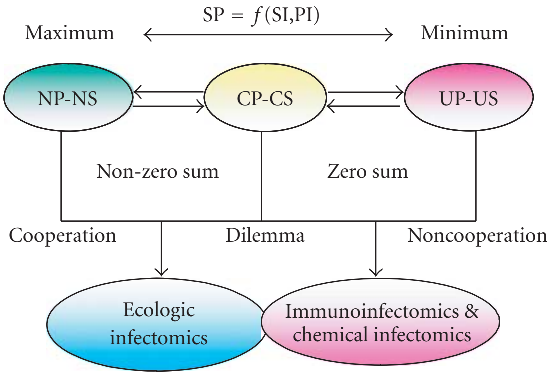


FIGURE 2: A continuum model of host-microbe interactions coupling with infectomic approaches to dissect the problems in microbial infections.

infectomic signatures specific for SI and PI. The set of symbiotic or pathogenic parameters is defined as a function SI($x$) or PI($x^*$). SI($x$) and PI($x^*$) are continuous functions ranging from 0 to 1 to admit different degrees of Sym and Pat, respectively. SI($x$) = 0 and PI($x^*$) = 0 indicate that $x$ and $x^*$ are perceived to be zero-symbiotic and zero-pathogenic, respectively. SI($x$) = 1 and PI($x^*$) = 1 indicate that $x$ and $x^*$ are perceived to be completely symbiotic and completely pathogenic, respectively. Intermediate values of SI($x$) and PI($x^*$) indicate that $x$ and $x^*$ are perceived to be partially symbiotic and partially pathogenic, respectively. Symbiotic points are used to determine the dynamic duality between Sym and Pat. SI and PI are interdependent parameters. The symbiotic point (SP) is a function of SI and PI:

$$\mathrm{SP} = f(\mathrm{SI}, \mathrm{PI}). \tag{1}$$

The focus of the dynamic duality research is to examine the ability of SP to transform situations of potential conflict (UP-US and CP-CS) into situations of cooperation (NP-NS). SP bears analogy to Schelling's focal point, which is any feature of such a game that provides a focus of convergence [16]. In the games with multiple Nash equilibria, one equilibrium usually stands out from the others (salient). Such an equilibrium is a focal point which can be easily recognized by all the players [12]. Thomas Schelling's *Strategy of Conflict* (1960) has been recognized as one of the most important works of game theory [11, 17]. There is no doubt that focal points play a central role in Schelling's game theory. Schelling has made a significant contribution to a reorientation of game theory. Understanding focal points is not only a key to improving game theory but also a key to dissecting SPs.

### *2.3. Game theoretical models (GTMs) of microbial infections*

In this paper, three types of GTMs are proposed for studies on NP-NS interactions (cooperative game), UP-US interactions (noncooperative games), and CP-CS interactions (dilemma or bargaining game). First, the NS-NS interactions are dissected with pure cooperative games in which each player chooses the strategy corresponding with the focal point in the expectation that the others will do the same. The significance of focal points can be shown most clearly in the pure cooperative games. As there is no conflict of interests in these games, all the players merely want to cooperate and they do not choose the alternative ways. Analysis of the cooperative game issues is to focus on coalition formation and distribution of the gains through cooperation. The SP in the NP-NS games tends to be maximal (see Figure 2). Secondly, noncooperative GTMs are used for analysis of the UP-US interactions. In contrast to cooperative games which focus on collective rationality and common interest, noncooperative games emphasize individual rationality and individual optimal strategy. The SP in the UP-US games tends to be minimal (see Figure 2). In games of pure conflict, defection is the equilibrium strategy and the total benefit to all players in the game, for every combination of strategies, always adds to zero (zero-sum). In the antagonistic UP-US interaction model, the surviving strategies of the UP community conflict with that of the US host. The UP evolves to exploit the host as much as possible, and the host adapts to exclude or limit the damage caused by the UP. Thirdly, we consider the strategic use of focal point theory in mixed situations to analyze the CP-CS interactions in which there is both conflict and mutual dependence. The most well-known example is the Prisoner's Dilemma game (a two-player game) in which each player chooses between a cooperating and defecting strategy. In this game, each player receives a higher playoff by defecting than by cooperating. However, a higher playoff is received if both cooperate than both defect. The two-player game can be extended to the N-player Prisoner's Dilemma game with arbitrary numbers of players.

## 3. RESULTS

### *3.1. Three communities in Escherichia coli species*

*E. coli* is one of the best understood and most thoroughly studied organisms and is advantageous as a model microorganism for the current studies. This bacterium is genotypically and phenotypically a highly diverse species, which is present in the three microbial communities (see Table 1). Most *E. coli* strains are commensals of higher vertebrates belonging to NP, but some are pathogenic (CP and UP). Uropathogenic *E. coli* (UPEC) in the CP group are the most common cause of community-acquired urinary tract infection (UTI). UPEC are responsible for about 80% of the estimated 150 million UTIs diagnosed annually [18]. *E. coli* O157, which belong to the UP group, is a major food pathogen. Shigella species, the cause of dysentery, are now known to be multiple distinct lineages of *E. coli*. Genomes of those *E. coli* strains have been sequenced (see Table 1). Recently, "better" *E. coli* strains (MDS41, 42, 43) have been engineered in which about 15% of the genome has been removed with the use of synthetic biology [19]. Coliphage, a virus which infects *E. coli*, is a major contributor responsible for diversification of *E. coli* [20, 21]. From a population-dynamic view, the interactions between coliphage and *E. coli* are analogous to those of a predator and a prey.

Table 1: *E. coli* in the three microbial communities.

| | *E. coli* strains | characteristics | Genome (Mb) | Putative<br>SI | <br>PI |
|---|---|---|---|---|---|
| NP | MG1655 (K12) | Commensal | 4.6 | >0.75 | <0.25 |
| | Nissle 1917 | Probiotics | 5.1 | >0.75 | <0.25 |
| | A0 34/86 | Probiotics | 4.8 | >0.75 | <0.25 |
| | MDS41, 42, 43 | K12 strains | 3.9 | >0.75 | <0.25 |
| CP | RS218 | Low pathogenicity | 5.1 | 0.5 ± 0.25 | 0.5 ± 0.25 |
| | CFT073 | Uropathogenicity | 5.2 | 0.5 ± 0.25 | 0.5 ± 0.25 |
| UP | O157 RIMD | High pathogenicity | 5.5 | <0.25 | >0.75 |
| | O157 EDL | High pathogenicity | 5.5 | <0.25 | >0.75 |
| | Shigella Sd197 | High pathogenicity | 4.4 | <0.25 | >0.75 |

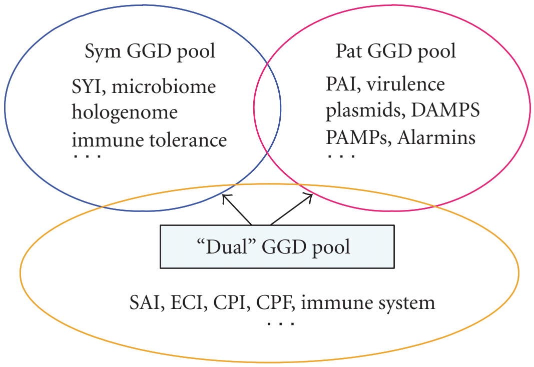


Figure 3: Gene pools contributing to the Sym-Pat duality. CPF: CP factors; CPI: CP islands; ECI: ecological islands; HPI: high pathogenicity islands; PAI: pathogenicity islands; SYI: symbiosis island.

### 3.2. *Sym-Pat duality is encoded by the genomes of both microbes and their hosts*

The development of microbial infections depends on the dynamic Sym-Pat duality, which is governed by the genomes of both the microbes and their hosts [2]. Molecular evolution of genetic structures for the Sym-Pat duality is influenced by both biotic and abiotic environmental factors in the ecosystems. There are three types of genetic/genomic determinants (GGDs) that may contribute to the Sym-Pat duality of microbial infections under specific environmental conditions (see Figure 3). The first type is the Sym GGD pool, which contributes to symbiosis. The Sym GGDs from the microbial partner include symbiosis-related genomic islands (SYI), plasmids, transposons, and microbiome, which is a collective genome of microbiota [2, 22, 23]. The gene pool contributing to microbial tolerance belongs to the host Sym GGDs [24]. The symbiotic homeostasis of the superorganism formed by the microbiota and its host is governed by hologenome, a complex of the host genome and microbiome [22, 25]. The second pool of GGDs contributes to the pathogenesis of microbial infections. These include pathogenicity islands (PAI), virulence plasmids, pathogen-associated molecular patterns (PAMPs), and endogenous alarmins [26]. PAMPs are a diverse group of microbial molecules, which are recognized by the host innate and adaptive immune system, primarily through toll-like receptors (TLRs) [26]. Alarmins are endogenous molecules within the host that signal tissue and cell damage. Effector cells of the innate and adaptive immunity can release alarmins when they are activated by PAMPs. Endogenous alarmins and exogenous PAMPs elicit similar responses by conveying a similar message. Therefore, they constitute a larger family of damage-associated molecular patterns (DAMPs). The third type of GGD pool has dual functions that depend on external and internal environments. These include ecological islands (ECI), certain GGDs from conditional pathogens (such as CP factors (CPFs) and CP islands (CPIs)), and the host GGDs with dual effects on microbial infections. The dual GGDs contribute to the Sym and Pat duality in specific ecological niches and within particular organisms. The same GGD may act as an SYI when the microbial recipient establishes a symbiotic relationship with its host, but becomes a PAI when it is adapting the pathogenic niche. A comparative infectomic study suggests that GimA, a 20-kb genomic island, is a typical CPI [27]. The dual biological functions of GimA depend on the genomic environments in *E. coli* strains. GimA present in meningitic *E. coli* K1 genome (O18:K1:H7) is essential for bacterial crossing the blood-brain barrier to cause meningitis [28]. In contrast, GimA is required for the probiotic function of *E. coli* K24 strain A0 34/86 (O83:K24:H31), which has been safely and effectively used in Czech pediatric clinics since 1967 [29]. The dual Sym-Pat properties of microbial determinants were also observed in *photorhabdus*, which is a genus of Gram-negative bacteria mutualistically associated with entomophagous nematodes of the family heterorhabditiae [30]. A hexA homologous gene from *photorhabdus* is able to regulate both symbiosis and pathogenesis [30]. Some microbes exhibit dual behavior as symbionts and pathogens in a manner dependent on the hosts. Sooty mangabey (SM) monkeys infected with simian immunodeficiency virus (SIV) do not develop acquired immunodeficiency syndrome (AIDS) [31]. In contrast, SIV infection of non-natural host monkeys, such as rhesus macaques (RMs), causes AIDS that closely resembles the human disease [31]. Similarly, polydnaviridae, a family of double-stranded DNA viruses, have evolved complex life cycles in which they interact as symbionts with one host and as pathogens with another [32]. All multicellular organisms, including human and flies, have evolved the conservative innate immune system as a double-edged sword [33]. It enables the host not only to combat pathogens but also to develop microbial tolerance to cohabit nonpathogenic microbiota by maintaining the homeostatic balance between the host and microorganisms [24, 33]. TLRs play central roles in the activating process of the innate immune system with dual functions. They have recently been shown to be involved in modulating intestinal homeostasis by recognizing commensal bacteria. They also sense extracellular PAMPs by triggering signaling, which results in the activation of proinflammatory (PI) pathways [24]. PI ligands of TLRs may be important for the activation and expansion of natural T regulatory cells (NatTReg), which control both deleterious and protective immune responses upon microbial infections [34]. Both innate TLRs and specific T cell receptors (TCR) contribute to the dual functions of NatTReg [34]. The full range of the dual GGDs in the immune system is unknown so far. However, it can be expected that the molecular evolution of the dual GGDs during the host-microbe coevolution will certainly lead to dynamic changes in the Sym-Pat duality.

### 3.3. *Duality relationship between Sym and Pat*

Sym and Pat, the fundamental components of microbial infection, can be defined as a function SI($x$) or PI($x^*$). SI($x$) and PI($x^*$) are continuous functions ranging from 0 to 1 to admit different degrees of Sym and Pat, respectively (see Figure 4). SI($x$) = 0 and PI($x^*$) = 0 indicate that $x$ and $x^*$ are perceived to be zero-symbiotic and zero-pathogenic, respectively. SI($x$) = 1 and PI($x^*$) = 1 indicate that $x$ and $x^*$ are perceived to be completely symbiotic and completely pathogenic, respectively. If SI is close to or equal to 1, microbial infection is a physiological process. It is now well accepted that mitochondria were derived from an endosymbiotic relationship with internalized proteobacteria, via a progressive transfer of genetic material [35]. This long symbiotic relationship reaches the maximum Sym value. The symbiotic nitrogen fixation process for converting atmospheric dinitrogen ($N_2$) to ammonia ($NH_3$) is essentially dependent on two partners: the host legume plant and bacteria belonging to the family *Rhizobiaceae* [36]. This type of microbial infection is more typically associated with a physiological process.

If the PI value approaches or reaches the maximum limit, microbial infection is more completely associated with a pathogenic process. Most of serious infectious diseases fall into this category. Intermediate values of SI($x$) and PI($x^*$) indicate that $x$ and $x^*$ are perceived to be partially symbiotic and partially pathogenic, respectively. Microbial infections induced by conditional pathogens represent a competitive relationship (see II or III in Figure 4). The hosts have large influences on SI and PI. For example, polydnaviruses have evolved complex life cycles in which they are able to adapt to a mutualistic partnership with one host and become pathogens with another. Their genomes reflect the dual roles as mutualists and pathogens [32].

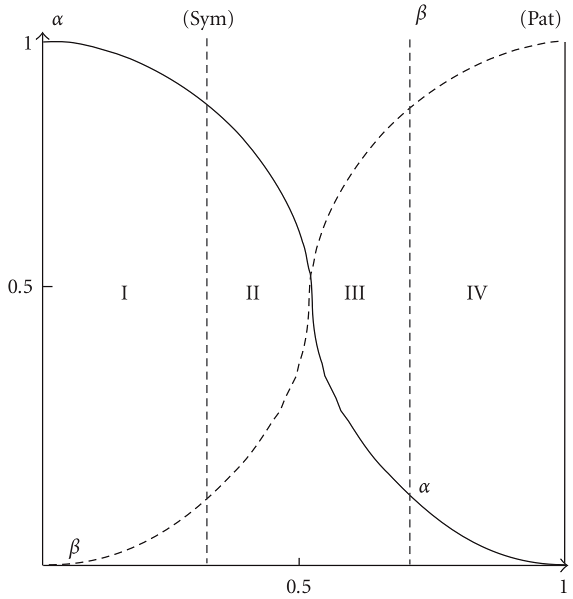
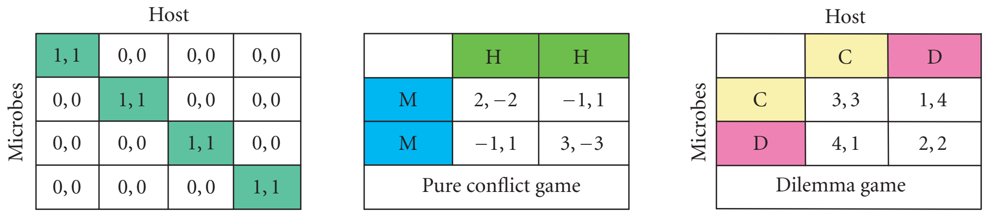


Figure 4: The duality relationship between Sym ($\alpha_i$) and Pat ($\beta_i$). A cooperative relationship (NP-NS: health and mutualism) (I) occurs between the host and the NP microbial community. A competitive relationship (CP-CS) (II and III) exists between the host and the CP microbial community. There are two types of competitions: better (II) and worse (III). An antagonistic relationship (UP-US) (IV) occurs between the host and the UP microbial community.

### 3.4. *Outcomes of three types of games in microbial infections*

The outcome of a game is not only determined by one individual's choices, but also depends on the strategies used by all the others. The dynamic Sym-Pat duality influenced by both microbes and their hosts is the key factor that determines the outcome of a game associated with microbial infection. One of the most fundamental issues in game theoretical solution concepts is that strategies used by individual players are based on the differences in payoff perceived by them [37]. This issue can be solved by Schelling's focal point theory. FPs constitute shared expectations that coordinate the activities of diverse players collectively or independently seeking their goals [38]. By harmonizing anticipated behaviors or responses despite the presence of imperfect information, individuals are able to coordinate their activities towards their ends. In 1960, Schelling classified games into three major categories: pure cooperative game on one side, pure conflict games on the other, and combinations of partial cooperation/partial conflict games in between [11]. Recently, a similar classification was designated in Gao's book, "Principles of Systemics" [39]. The pure cooperative game models are used for dissecting NP-NS interaction problems. The payoff matrix for a pure NP-NS problem would resemble something like that in Figure 5(a). In contrast, the pure UP-US interaction, a situation of pure antagonism, is characterized by completely opposing interests, where the pathogenesis of microbial infection is the most predominant event. The payoff matrix for the pure UP-US game would look something like that



Host

| Microbes | | | |
|---|---|---|---|
| 1, 1 | 0, 0 | 0, 0 | 0, 0 |
| 0, 0 | 1, 1 | 0, 0 | 0, 0 |
| 0, 0 | 0, 0 | 1, 1 | 0, 0 |
| 0, 0 | 0, 0 | 0, 0 | 1, 1 |

(a)

| | H | H |
|---|---|---|
| M | 2, −2 | −1, 1 |
| M | −1, 1 | 3, −3 |

Pure conflict game

(b)

Host

| Microbes | C | D |
|---|---|---|
| C | 3, 3 | 1, 4 |
| D | 4, 1 | 2, 2 |

Dilemma game

(c)

Figure 5: (a) Pure cooperative game. (b) Pure conflict game. H: host; M: microbes. (c) Dilemma game. The number in the left of each pair indicates the payoff for microbes, and the right, the host. Higher numbers represent greater payoff for the individual. Two strategies (Cooperation (C) and Defection (D)) are used.

in Figure 5(b). The state of microbial infection has conventionally been characterized as lying on the extreme conflict end. This depiction comes from the Manichaean view of the microbe-human host relationship. This situation is depicted in Figure 5(b) as a two-player game. In the pure conflict game, the relationships between the microbes and their hosts end up in a "war of all against all" in which the payoffs of the outcomes add to zero (zero-sum). The three microbial community principle and the dynamic Sym-Pat duality concept would help establish a holistic view of microbial infections. The CP-CS interaction is a situation where cooperation and conflict coexist. As illustrated in Figure 5(c), the payoffs for the CP-CS games would lie midway between the pure cooperative and pure conflict games. The CP-CS problems are microbial dilemmas, in which there is a mixture of mutual dependence and conflict of partnerships and competition. The underlying idea arises naturally from the well-known games for the social dilemmas and the Prisoner's Dilemma (PD), in which each player chooses between a cooperating and a defecting strategy [40]. As shown in Figure 5(c), each player receives a higher playoff by defecting than by cooperating, no matter what the other player chooses. However, they receive a higher playoff if both cooperate than both defect. The CP-CS interactions can coevolve toward two different directions, increasing (more cooperation) or decreasing (more antagonism) the SP (see Figure 2).

## 4. DISCUSSION

In this paper, focal point theory-based game models are proposed for analysis of the dynamic Sym-Pat duality in microbial infections. DSPD is the most fundamental problem in infectomics, which is the integration of omics and mathematical/computational approaches. There are three types of infectomic approaches that can be used for the control of microbial infections: ecological infectomics, immunoinfectomics, and chemoinfectomics [2]. Ecological infectomics will explore symbiotic solutions to microbial infections. Developing novel immunological intervention strategies for the prevention and treatment of microbial infections using infectomic signatures and immunomic approaches falls within the field of immunoinfectomics. Chemoinfectomics represents the most powerful approach to the development of a new generation of drugs for antimicrobial chemotherapy.

### 4.1. *Symbiosis point converting (SPC): ecological infectomics-based approaches for rational control of microbial infections*

As microbial infection is an ecological and evolutionary paradigm which is associated with coevolution between hosts and microbes (such as human host and microorganisms, phages, and bacteria) in dynamic ecosystems, two ecological infectomics-based SPC approaches (increasing and decreasing SP) can be used for rational control of infectious diseases [2]. The focus in SP increasing approaches is how to transform situations of potential conflict (pathogenesis) into cooperation (symbiosis) by dissecting the dynamic duality relationships between Sym and Pat in microbial infections and developing symbiotic agents (symbiotics) that favor a healthy symbiosis [2]. Symbiotics are defined as products that are beneficial to symbiotic ecology of the superorganisms consisting of microbes and their human hosts. These include microbial (e.g., probiotic bacteria) and nonmicrobial agents (e.g., prebiotics) [2]. The introduction of beneficial symbiotics with higher SP in our body should be a very attractive rationale for modulating the microbiota, improving the symbiotic homeostasis of the superorganism, and providing a microbial stimulus to the host immune system against pathogens. The use of probiotics has been suggested as a promising approach for combating infectious diseases, and delivering drugs and vaccines [2]. Decreasing SP is another rational strategy for control of microbial infections. As phages, which specifically kill bacteria, play an important role in the ecology, evolution, and virulence of a number of pathogens, there is a rational use of phages for treatment and prevention of bacterial infections. The use of phages to treat bacterial infections has a long history dating back to mid 1910's [2]. Due to the availability of effective broad-spectrum antibiotics in the early 1940's, phage therapy was discarded in Western medicine at that time. The rising crisis of antibiotic resistance has recently increased great interest in phages and their use as natural antimicrobial agents to fight microbial infections [2]. Compared with commonly used antibiotics, a great advantage of phages is their narrow host range. Recent studies have shown that coinfection with GB virus C (GBV-C) is associated with a decreased mortality in HIV-infected patients [41]. Therefore, reducing SP between microbial agents (such as phages and GBV-C) and targeted pathogens is another excellent ecological approach for the development of novel antimicrobial agents.

### 4.2. *Specific pathogen-targeting (SPT): immunoinfectomics- and chemoinfectomics-based approaches for prevention and treatment of infectious diseases*

In contrast to the ecological infectomics-based SPC approaches that focus on the symbiotic relationships (such as NP-NS and CP-CS interactions) between the hosts and microbial communities, immunoinfectomics- and chemoinfectomics-based SPT approaches emphasize the use of antagonistic relationships (such as UP-US interactions) between the hosts and microorganisms. It is important to point out that the SPT approaches are intrinsically different from the conventional pathogen-targeting antimicrobial agents, which kill both pathogens and nonpathogens [2]. The availability of the genomic information from both microbes and their hosts has resulted in exciting new progress in the field of immunoinfectomics. Nanobody (the smallest fragment of naturally occurring single-domain antibody)-based technologies and immune epitope mapping have emerged as the very powerful tools for the discovery and development of novel antimicrobial agents [2]. Recently, a nanobody-conjugated human trypanolytic factor has been successfully used for an experimental therapy of African trypanosomiasis [42]. Concurrent advances in both high-throughput chemistry and infectomics have given rise to the field of chemoinfectomics for elucidating and validating drug targets, and generating novel therapeutics. Chemoinfectomics refer to the use of small synthetic molecules that are highly specific for defined infectomic targets, for biological function analysis and to discover new drug leads. The progress towards understanding the dynamic Sym-Pat duality in microbial infections using focal point theory-based game models will greatly facilitate the use of ecological infectomics, immunoinfectomics, and chemoinfectomics for the rational control of infectious diseases.

## REFERENCES

[1] S.-H. Huang, T. Triche, and A. Jong, “Infectomics: genomics and proteomics of microbial infections,” *Functional & Integrative Genomics*, vol. 1, no. 6, pp. 331–344, 2002.

[2] S.-H. Huang, X. Wang, and A. Jong, “The evolving role of infectomics in drug discovery,” *Expert Opinion on Drug Discovery*, vol. 2, no. 7, pp. 961–975, 2007.

[3] M. J. Roossinck, “Symbiosis versus competition in plant virus evolution,” *Nature Reviews Microbiology*, vol. 3, no. 12, pp. 917–924, 2005.

[4] G. G. Dimijian, “Evolving together: the biology of symbiosis,” *Baylor University Medical Center Proceedings*, vol. 13, no. 3, pp. 217–226, 2000.

[5] L. Smolin, *Three Roads to Quantum Gravity*, Basic Books, New York, NY, USA, 2001.

[6] J. Lederberg, “Infectious history,” *Science*, vol. 288, no. 5464, pp. 287–293, 2000.

[7] J. Lederberg, “The future of infectious diseases,” *Journal of Urban Health*, vol. 75, no. 3, pp. 463–470, 1998.

[8] H. Brüssow, C. Canchaya, and W.-D. Hardt, “Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion,” *Microbiology and Molecular Biology Reviews*, vol. 68, no. 3, pp. 560–602, 2004.

[9] M. E. J. Woolhouse, J. P. Webster, E. Domingo, B. Charlesworth, and B. R. Levin, “Biological and biomedical implications of the co-evolution of pathogens and their hosts,” *Nature Genetics*, vol. 32, no. 4, pp. 569–577, 2002.

[10] A. Casadevall and L.-A. Pirofski, “Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease,” *Infection and Immunity*, vol. 68, no. 12, pp. 6511–6518, 2000.

[11] T. C. Schelling, *The Strategy of Conflict*, Harvard University Press, Cambridge, Mass, USA, 1960.

[12] J. Mehta, C. Starmer, and R. Sugden, “Focal points in pure coordination games: an experimental investigation,” *Theory and Decision*, vol. 36, no. 2, pp. 163–185, 1994.

[13] R. E. Ley, D. A. Peterson, and J. I. Gordon, “Ecological and evolutionary forces shaping microbial diversity in the human intestine,” *Cell*, vol. 124, no. 4, pp. 837–848, 2006.

[14] S. Falkow, “Is persistent bacterial infection good for your health?” *Cell*, vol. 124, no. 4, pp. 699–702, 2006.

[15] D. Kuklinska and M. Kilian, “Relative proportions of *Haemophilus* species in the throat of healthy children and adults,” *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 3, no. 3, pp. 249–252, 1984.

[16] K. Binmore and L. Samuelson, “The evolution of focal points,” *Games and Economic Behavior*, vol. 55, no. 1, pp. 21–42, 2006.

[17] R. Sugden and I. E. Zamarrón, “Finding the key: the riddle of focal points,” *Journal of Economic Psychology*, vol. 27, no. 5, pp. 609–621, 2006.

[18] R. A. Welch, V. Burland, G. Plunkett III, et al., “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 17020–17024, 2002.

[19] G. Pósfai, G. Plunkett III, T. Fehér, et al., “Emergent properties of reduced-genome *Escherichia coli*,” *Science*, vol. 312, no. 5776, pp. 1044–1046, 2006.

[20] H. Brüssow, “Phage therapy: the *Escherichia coli* experience,” *Microbiology*, vol. 151, no. 7, pp. 2133–2140, 2005.

[21] M. Ohnishi, K. Kurokawa, and T. Hayashi, “Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?” *Trends in Microbiology*, vol. 9, no. 10, pp. 481–485, 2001.

[22] D. A. Relman, “New technologies, human-microbe interactions, and the search for previously unrecognized pathogens,” *Journal of Infectious Diseases*, vol. 186, supplement 2, pp. S254–S258, 2002.

[23] J. Hacker and E. Carniel, “Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes,” *EMBO Reports*, vol. 2, no. 5, pp. 376–381, 2001.

[24] P. J. Sansonetti, “War and peace at mucosal surfaces,” *Nature Reviews Immunology*, vol. 4, no. 12, pp. 953–964, 2004.

[25] E. Rosenberg, O. Koren, L. Reshef, R. Efrony, and I. Zilber-Rosenberg, “The role of microorganisms in coral health, disease and evolution,” *Nature Reviews Microbiology*, vol. 5, no. 5, pp. 355–362, 2007.

[26] M. E. Bianchi, “DAMPs, PAMPs and alarmins: all we need to know about danger,” *Journal of Leukocyte Biology*, vol. 81, no. 1, pp. 1–5, 2007.

[27] S.-H. Huang, "Infectomics-based new theory for dissecting the pathogenesis, prevention and therapeutics of microbial infections," in *Teda-Watson International forum on Biotechnology and Biomedicine*, vol. 4, pp. 73–74, Tianjin, China, 2005.

[28] S.-H. Huang, Y.-H. Chen, G. Kong, et al., "A novel genetic island of meningitic *Escherichia coli* K1 containing the *ibeA* invasion gene (GimA): functional annotation and carbon-source-regulated invasion of human brain microvascular endothelial cells," *Functional & Integrative Genomics*, vol. 1, no. 5, pp. 312–322, 2001.

[29] J. Hejnova, U. Dobrindt, R. Nemcova, et al., "Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83:K24:H31)," *Microbiology*, vol. 151, no. 2, pp. 385–398, 2005.

[30] S. A. Joyce and D. J. Clarke, "A *hexA* homologue from *Photorhabdus* regulates pathogenicity, symbiosis and phenotypic variation," *Molecular Microbiology*, vol. 47, no. 5, pp. 1445–1457, 2003.

[31] G. Silvestri, "Naturally SIV-infected sooty mangabeys: are we closer to understanding why they do not develop AIDS?" *Journal of Medical Primatology*, vol. 34, no. 5-6, pp. 243–252, 2005.

[32] B. A. Webb, M. R. Strand, S. E. Dickey, et al., "Polydnavirus genomes reflect their dual roles as mutualists and pathogens," *Virology*, vol. 347, no. 1, pp. 160–174, 2006.

[33] K. S. Kobayashi and R. A. Flavell, "Shielding the double-edged sword: negative regulation of the innate immune system," *Journal of Leukocyte Biology*, vol. 75, no. 3, pp. 428–433, 2004.

[34] J. Demengeot, S. Zelenay, M. F. Moraes-Fontes, Í. Caramalho, and A. Coutinho, "Regulatory T cells in microbial infection," *Springer Seminars in Immunopathology*, vol. 28, no. 1, pp. 41–50, 2006.

[35] R. Lucattini, V. A. Likić, and T. Lithgow, "Bacterial proteins predisposed for targeting to mitochondria," *Molecular Biology and Evolution*, vol. 21, no. 4, pp. 652–658, 2004.

[36] T. Uchiumi, T. Ohwada, M. Itakura, et al., "Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome," *Journal of Bacteriology*, vol. 186, no. 8, pp. 2439–2448, 2004.

[37] M. C. W. Janssen, "Rationalizing focal points," *Theory and Decision*, vol. 50, no. 2, pp. 119–148, 2001.

[38] P. T. Leeson, C. J. Coyne, and P. J. Boettke, "Converting social conflict: focal points and the evolution of cooperation," *The Review of Austrian Economics*, vol. 19, no. 2-3, pp. 137–147, 2006.

[39] L. C. Gao, *Principles of Systemics*, Science Press, Beijing, China, 2005.

[40] A. M. Colman, "Thomas C. Schelling's psychological decision theory: introduction to a special issue," *Journal of Economic Psychology*, vol. 27, no. 5, pp. 603–608, 2006.

[41] D. E. Yirrell, E. Wright, L. A. Shafer, et al., "Association between active GB virus-C (hepatitis G) infection and HIV-1 disease in Uganda," *International Journal of STD & AIDS*, vol. 18, no. 4, pp. 244–249, 2007.

[42] T. N. Baral, S. Magez, B. Stijlemans, et al., "Experimental therapy of African trypanosomiasis with a nanobody-conjugated human trypanolytic factor," *Nature Medicine*, vol. 12, no. 5, pp. 580–584, 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 837653, 9 pages
doi:10.1155/2008/837653

*Research Article*

# Receptor Guided 3D-QSAR: A Useful Approach for Designing of IGF-1R Inhibitors

**M. Muddassar, F. A. Pasha, H. W. Chung, K. H. Yoo, C. H. Oh, and S. J. Cho**

*Future Fusion Technology Division, Computational Science Center, Korea Institute of Science and Technology,*
*P.O. Box 131, Cheongryang, Seoul 130-650, South Korea*

Correspondence should be addressed to S. J. Cho, chosj@kist.re.kr

Received 16 August 2007; Accepted 17 December 2007

Recommended by Daniel Howard

Research by other investigators has established that insulin-like growth factor-1 receptor (IGF-1R) is a key oncological target, and that derivatives of 1, 3-disubstituted-imidazo[1,5-$\alpha$] pyrazine are potent IGF-1R inhibitors. In this paper, we report on our three-dimensional quantitative structure activity relationship (3D-QSAR) studies for this series of compounds. We validated the 3D-QSAR models by the comparison of two major alignment schemes, namely, ligand-based (LB) and receptor-guided (RG) alignment schemes. The latter scheme yielded better 3D-QSAR models for both comparative molecular field analysis (CoMFA) ($q^2 = 0.53$, $r^2 = 0.95$) and comparative molecular similarity indices analysis (CoMSIA) ($q^2 = 0.51$, $r^2 = 0.86$). We submit that this might arise from the more accurate inhibitor alignment that results from using the structural information of the active site. We conclude that the receptor-guided 3D-QSAR may be helpful to design more potent IGF-1R inhibitors, as well as to understand their binding affinity with the receptor.

## 1. INTRODUCTION

The insulin-like growth factor-1 receptor is a membrane-associated receptor that belongs to subclass I of the receptor tyrosine kinase (RTK) superfamily [1]. IGF-1R has been shown to have significant roles in the regulation of normal cell growth. It has mitogenic and survival effects on human cancer cells [2]. The Binding of IGF-1 to IGF-1R activates the RTK, and later, in turn, activates a cascade of downstream signals, which are postulated to stimulate cell proliferation and enhance resistance to apoptosis [3]. Understandably, the abnormal expression of the IGF-1R has been implicated to cancer. Epidemiological studies have also shown a link between serum concentrations of IGF-1 and IGFBP-3 with increased risks of breast cancer [4]. A number of anticancer agents which inhibit the IGF-1R activity and proliferation [5] have been extracted from plants [6] as well as synthesized, such as BMS-554417 (2-(4-substituted-2-oxo-1,2-dihydropyridin-3-yl)-benzimidazole) [7] and NVP-AEW541 (pyrrolo[2,3-*d*] pyrimidine derivative) molecules. Both of these compounds are orally administered and have proved antitumor activity. Various QSAR techniques are being used to explore more potent ligands [8–11]; but in this study, we performed comparative three-dimensional quantitative structure activity relationship (3D-QSAR) [12–14] analyses on IFG-1R inhibitors [15] of imidazo [1, 5-$\alpha$] pyrazine derivatives. In 3D-QSAR [14], determination of the bioactive conformer [16] and molecular alignment of the compounds is key factor to get meaningful results. The biologically active conformations of the structures should be aligned in a way that represents a similar binding mode [17]. Here we first applied the ligand-based (LB) strategy using the systematic search-based minimum energy conformer approach [18]. Second, receptor-based 3D-QSAR [19] using molecular docking of inhibitors in the available X-ray crystal structure [20] of the receptor protein. The qualities of these 3D-QSAR models were compared and discussed with respect to the IGF-1R target.

## 2. MATERIAL AND METHODS

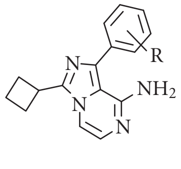
A series of 54 potent 1, 3-disubstituted imidazole [1, 5-$\alpha$] pyrazine derivatives with their inhibitory activities to IGF-1R were taken from the literature [15]. The dataset was randomly divided into 43 and 11 molecules, the training and test datasets, respectively. The observed $IC_{50}$ values were converted into $pIC_{50}$ values and are reported in Table 1.

Table 1: The structures and observed IGF inhibitory activities [15].

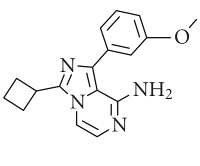
| No. | Structure | R | IC$_{50}$ (M) | pIC$_{50}$ |
|---|---|---|---|---|
| 1 | A | 4-OBn | $1.97 \times 10^{-6}$ | 5.706 |
| 2 | A | 3-OH | $0.518 \times 10^{-6}$ | 6.286 |
| 3 | A | 3-OBn–4-OMe | $1.35 \times 10^{-6}$ | 5.870 |
| 4 | A | 3-OBn–4-OH | $3.31 \times 10^{-6}$ | 5.480 |
| 5* | B | Cyclopentyl | $3.5 \times 10^{-6}$ | 5.456 |
| 6 | B | Cyclohexyl | $1.05 \times 10^{-6}$ | 5.979 |
| 7* | B | –CH$_2$–cyclopropyl | $2.27 \times 10^{-6}$ | 5.644 |
| 8* | B | –CH$_2$–cyclohexyl | $1.11 \times 10^{-6}$ | 5.955 |
| 9 | B | –CH$_2$CH$_2$OMe | $6.28 \times 10^{-6}$ | 5.202 |
| 10 | B | –CH$_2$–2-pyridyl | $1.09 \times 10^{-6}$ | 5.963 |
| 11 | C | H | $0.606 \times 10^{-6}$ | 6.218 |
| 12 | C | 2-F | $0.224 \times 10^{-6}$ | 6.650 |
| 13 | C | 3-F | $0.51 \times 10^{-6}$ | 6.292 |
| 14 | C | 4-F | $1.23 \times 10^{-6}$ | 5.910 |
| 15 | C | 2-Cl | $0.343 \times 10^{-6}$ | 6.465 |
| 16 | C | 3-Cl | $2.12 \times 10^{-6}$ | 5.674 |
| 17* | C | 4-Cl | $0.980 \times 10^{-6}$ | 6.009 |
| 18 | C | 2-OCF$_2$H | $3.28 \times 10^{-6}$ | 5.484 |
| 19 | C | 3-OCF$_2$H | $5.78 \times 10^{-6}$ | 5.238 |
| 20 | C | 4-OCF$_2$H | $2.82 \times 10^{-6}$ | 5.550 |
| 21 | C | 2,3-Difluoro | $0.898 \times 10^{-6}$ | 6.047 |
| 22 | C | 3,4-Difluoro | $4.48 \times 10^{-6}$ | 5.349 |
| 23* | C | 2,5-Difluoro | $0.329 \times 10^{-6}$ | 6.483 |
| 24 | C | 2,6-Difluoro | $0.215 \times 10^{-6}$ | 6.668 |
| 25 | C | 3,5-Difluoro | $1.35 \times 10^{-6}$ | 5.870 |
| 26 | C | 2,6-Dichloro | $1.67 \times 10^{-6}$ | 5.777 |
| 27 | C | 2-Cl,6-F | $0.248 \times 10^{-6}$ | 6.606 |
| 28 | D | Cyclopentyl | $1.05 \times 10^{-6}$ | 5.979 |
| 29 | D | Cyclohexyl | $3.51 \times 10^{-6}$ | 5.455 |
| 30 | D | Cycloheptyl | $3.79 \times 10^{-6}$ | 5.421 |
| 31* | D | Phenyl | $1.68 \times 10^{-6}$ | 5.775 |
| 32 | E | trans-NH$_2$ | $0.221 \times 10^{-6}$ | 6.656 |
| 33 | E | cis-NH$_2$ | $0.775 \times 10^{-6}$ | 6.111 |
| 34* | E | trans-NHMe | $0.105 \times 10^{-6}$ | 6.979 |
| 35 | E | trans-Pyrrolodinyl | $1.82 \times 10^{-6}$ | 5.740 |
| 36* | E | trans-Piperidinyl | $3.40 \times 10^{-6}$ | 5.469 |
| 37 | E | trans-NHPh | $1.30 \times 10^{-6}$ | 5.886 |
| 38 | E | trans-NHBn | $1.39 \times 10^{-6}$ | 5.857 |
| 39 | F | trans-NH$_2$ | $0.119 \times 10^{-6}$ | 6.924 |
| 40 | F | cis-NH$_2$ | $0.228 \times 10^{-6}$ | 6.642 |
| 41 | F | trans-N(Et)$_2$ | $0.115 \times 10^{-6}$ | 6.939 |
| 42 | F | trans-Azetidinyl | $0.081 \times 10^{-6}$ | 7.092 |
| 43 | F | trans-Pyrrolidinyl | $0.103 \times 10^{-6}$ | 6.987 |
| 44* | F | trans-Morpholino | $0.091 \times 10^{-6}$ | 7.041 |
| 45* | G | trans-Pyrrolidinyl | $0.116 \times 10^{-6}$ | 6.936 |
| 46 | G | cis-Pyrrolidinyl | $0.089 \times 10^{-6}$ | 7.051 |
| 47 | G | cis-NH$_2$ | $0.060 \times 10^{-6}$ | 7.222 |
| 48 | G | cis-NMe$_2$ | $0.166 \times 10^{-6}$ | 6.780 |
| 49* | G | cis-Piperidinyl | $0.237 \times 10^{-6}$ | 6.625 |
| 50 | G | cis-Morpholino | $0.148 \times 10^{-6}$ | 6.830 |
| 51 | G | cis-NH-iPr | $0.220 \times 10^{-6}$ | 6.658 |
| 52 | G | cis-N(Me)-Piperizinyl | $0.265 \times 10^{-6}$ | 6.577 |

Table 1: Continued.

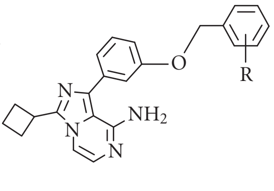
| No. | Structure | R | $IC_{50}$ (M) | $pIC_{50}$ |
|---|---|---|---|---|
| 53 | H | trans-$NH_2$ | $0.526 \times 10^{-6}$ | 6.279 |
| 54 | H | cis-$NH_2$ | $0.554 \times 10^{-6}$ | 6.256 |



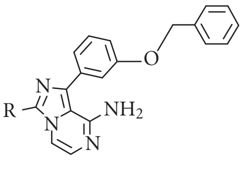
A



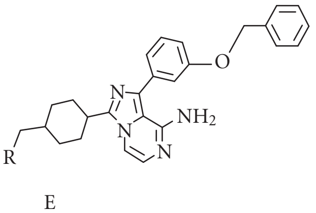
B



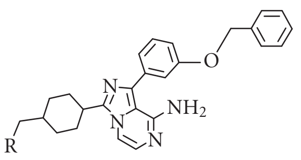
C



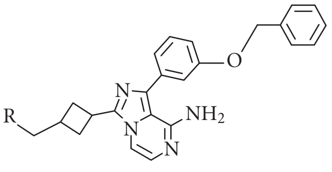
D



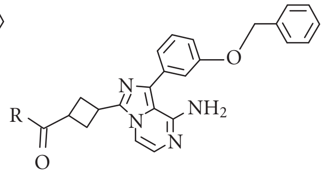
E



F



G



H

*Compounds used in the test set.

### 2.1. Computational details

The molecular modeling studies were carried out using SYBYL 7.3. The initial structures were minimized at Tripos force field [21] with MMFF94 charge by using conjugate gradient method, and convergence criterion was 0.005 kcal/mol.The comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) studies require aligned structures [16]. The ligand-based (LB) and receptor-guided (RG) alignment techniques were used in two geometrical schemes respectively.

### 2.2. CoMFA and CoMSIA

Lennard-Jones and Coulomb potentials-based CoMFA analysis has been performed and the steric as well as electrostatic energies were calculated by using $sp^3$ carbon probe atom with Van der Waal radius of 1.52 Å and +1 charge. The energies were truncated to ±30 kcal/mol and the electrostatic contributions were ignored at lattice interactions with maximum steric interactions. The CoMFA were generated by standard method in SYBYL. The CoMSIA models were also derived with the same lattice box used as in CoMFA calculations. All five CoMSIA similarity indices (steric, electrostatic, hydrophobic, H-bond donor, and H-bond acceptor) were evaluated using the probe atom. The CoMSIA models from hydrophobic and H-bonds were calculated between the grid point and each atom of the molecule by a Gaussian function [14]. An attenuation factor's default value of 0.30 was used, which is the standard distance dependence of molecular similarity.

### 2.3. PLS analysis and validation of QSAR models

In orderto derive 3D-QSAR models, the CoMFA and CoMSIA descriptors were used as independent variables and the $pIC_{50}$ values as the dependent variable. Partial least-square (PLS) method [22] was used to linearly correlate these CoMFA and CoMSIA descriptors to the inhibitory activity values. The CoMFA cutoff values were set to 30 kcal/mol for both steric and electrostatic fields, and also all fields were scaled by the default options in SYBYL. The cross-validation analysis was performed using the leave one out (LOO) method in which one compound is removed from the dataset and its activity is predicted using the model derived from the rest of the dataset. The cross-validated correlation coefficient ($q^2$) that resulted in optimum number of components and lowest standard error of prediction were calculated using the following formulae,

$$q^2 = 1 - \frac{\sum_y \left(y_{\text{pred}} - y_{\text{observed}}\right)^2}{\sum_y \left(y_{\text{observed}} - y_{\text{mean}}\right)^2}, \qquad \text{PRESS} = \sum_y \left(y_{\text{predicted}} - y_{\text{observed}}\right)^2, \tag{1}$$

where $y_{\text{pred}}$, $y_{\text{actual}}$, and $y_{\text{mean}}$ are predicted, actual, and mean values of the target property ($pIC_{50}$), respectively. The non-cross-validated PLS analyses were performed with column filtering value of 2.0, to reduce analysis time with small effect on the $q^2$ values. To further assess the robustness and statistical confidence of the derived models, bootstrapping analysis for 100 runs were performed.

The predictive power of 3D-QSAR models, derived by using the training set were examined by an external test set of eleven molecules. The predictive ability of the models is expressed by the predictive $r^2$ value, which is analogous to cross-validated $r^2$ ($q^2$) and is calculated using the following formula:

$$r^2_{\text{pred}} = \frac{\text{SD} - \text{PRESS}}{\text{SD}}, \tag{2}$$

where SD is the sum of the squared deviations between the biological activities of the test set and mean activities of the training molecules and PRESS is the sum of squared deviation between predicted and actual activities of the test set molecules.

## 3. RESULTS AND DISCUSSION

### 3.1. Ligand-based alignment

In this scheme, the most active molecule was used as a template. Systematic search routine was used in the conformational analysis and all rotatable bonds were searched in 10° increments from 0° to 350°. Conformational energies were computed with electrostatic term, and the lowest energy conformer was selected. The template was modified for other ligands of the series. All ligands were minimized by Tripos force field but the common moiety was constrained during minimization. The molecules were aligned by superimposing common substructures using SYBYL database alignment option. These aligned structures were subsequently used for ligand-based CoMFA/CoMSIA probe interaction energy calculations.

### 3.2. Receptor-guided alignment

This geometrical scheme is based on docked geometry. The best docked mode of the smallest compound was taken as template and modified for the other compounds. The compounds were minimized by Tripos force field (Powell method, 2000 iterations, and $0.05\,\text{kcal}\cdot\text{mol}^{-1}\cdot\text{Å}^{-1}$ energy gradient convergence criteria). All minimized structures at this binding mode were superimposed to get the molecular alignment for CoMFA and CoMSIA. The superimposed structures inside the receptor site were further used for CoMFA and CoMSIA analysis.

### 3.3. Molecular docking

The structure coordinates of IGF-1R were obtained from protein databank (1JQH) [20]. Recently, Mulvihill et al. [15] presented a possible binding mode of compound-2 by using FlexX-based docking. Here we have also performed molecular docking of same compound. The PDB file obtained from protein data bank was used as receptor site. All water molecules were removed and the protein was modified to dock inhibitor. The active site was defined with a distance of 6.5 Å of ATP binding site. The ligand-2 was docked into the monomer unit (A) of IGF-1R and out of 100 conformers the best mode was selected as template. This binding mode seems prominent as the hydrophobic zone of inhibitor corresponds to hydrophobic pocket of IGFR. The residue E-1080, M-1082, K-1033, D-1086, G-1006, and L-1005 makes hinge contact and might have significant role in the inhibition of IGF-1R. It is also clear from all the figures that the depicted mode holds 3 H-bonds in this region. The –OH group of benzene ring makes H-bond with –NH of K-1033, nitrogen of pyrimidine ring makes contact with –NH of M-1082 and both act as H-bond acceptor. The $NH_2$ group of pyrimidine ring acts as H-bond donor and makes contact with oxygen of E1080.

### 3.4. CoMFA and CoMSIA results

The CoMFA and CoMSIA studies were carried out by using both geometrical schemes with different descriptors fields independently and in combination. The ligand-based alignment gave better results for CoMFA model using both field descriptors with cross-validated $r^2(q^2) = 0.52$ and non-cross-validated $r^2 = 0.88$, while for CoMSIA model, combination of steric, electrostatic, and H-bond acceptor yielded the best statistical values with $q^2 = 0.42$ and $r^2 = 0.80$. The internal predictivity of these CoMFA and CoMSIA models was also good with boot-strapped correlation coefficient $r^2_{\text{bs}} = 0.91$ and 0.85, respectively. These models were also validated on a test set of 11 molecules with predictive $r^2 = 0.67$ for CoMFA model and 0.57 for CoMSIA model. In comparison to LB, receptor-guided alignment yielded more significant models with better understanding of these inhibitors and receptor interactions. Best CoMFA models were obtained by combination of steric and electrostatic field descriptors with $q^2 = 0.53$ and $r^2 = 0.95$. Whereas steric, electrostatic, and H-bond acceptor filed descriptors gave the best CoMSIA model with $q^2 = 0.51$ and $r^2 = 0.86$. To further asses the robustness and statistical confidence, the boot strapping analysis were performed for 100 runs. The $r^2_{\text{bs}}$ for CoMFA = 0.97 and CoMSIA = 0.90 models suggest that a good internal consistency exists within the underlying dataset. The high $r^2$ predictive values for CoMFA and CoMSIA (0.67 and 0.64, resp.) also prove models validity. In our efforts to obtain the more pronounced model, region focusing was performed. It only yielded high $q^2$ value which is not sufficient condition for the model to have high predictive power [23]. The regression summary of different 3D-QSAR models obtained at default parameters and after region focusing are presented in Tables 2 and 3, respectively. The predicted $\text{pIC}_{50}$ values for training and test set from CoMFA and CoMSIA models are given in Tables 4 and 5, respectively.

In 3D-QSAR, the determination of the bioactive conformer and molecular alignment of the compounds is an important step. In ligand-based techniques, the minimum energy conformers are often used as bioactive conformer. In contrast, the binding poses obtained from cocrystal structure are used in receptor-guided techniques. Here, both techniques were used. The statistical results indicate that conformation obtained from molecular docking is more reliable. In Figure 1, the yellow conformer displays systematic search-based minimum energy conformer while the red structure shows docked conformer. The findings are reasonable as the oxygen attached with benzyl group of docked conformer is more closed to amino acid (Asp1086) that facilitates an H-bonding between –NH of Asp-1086 and this oxygen atom of the inhibitor; but in case of minimum energy conformer (yellow), the benzyl moiety is quite far and disfavors such interactions.

### 3.5. The CoMFA contour maps

Figures 2 and 3 show the electrostatic and steric contour maps of the best models based on receptor-guided alignment scheme. The electrostatic interactions are represented by

Table 2: Statistical summary of different PLS analysis. (GS: geometrical scheme; SE: standard error of estimate; n.: number of components; $F$: Fischer's $F$ value for test of significance; $r^2_{\mathrm{bs}}$: coefficient of determination after 100 bootstrapping runs; SD: standard deviation; Field contribution: (S) steric field, (E) electrostatic field, (H) hydrophobic field, (D) H-bond donor field, and (A) H-bond acceptor field.).

| Analysis | GS | Field | $q^2$ | n. | $r^2$ | $F$ | SE | $r^2_{\mathrm{bs}}$ | SD | $r^2_{\mathrm{pred}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| CoMFA | LB | S | 0.52 | 4 | 0.84 | 50.12 | 0.23 | — | — | — |
| CoMFA | LB | E | 0.38 | 3 | 0.73 | 34.7 | 0.30 | — | — | — |
| CoMFA | LB | 0.49S/0.51E | 0.52 | 4 | 0.88 | 70 | 0.20 | 0.91 | 0.1 | 0.67 |
| CoMFA | RG | S | 0.38 | 4 | 0.72 | 24.5 | 0.31 | — | — | — |
| CoMFA | RG | E | 0.42 | 7 | 0.86 | 33.12 | 0.22 | — | — | — |
| CoMFA | RG | 0.45/0.55 | 0.53 | 6 | 0.95 | 113.6 | 0.13 | 0.97 | 0.00 | 0.67 |
| CoMSIA | LB | S | 0.27 | 3 | — | — | — | — | — | — |
| CoMSIA | LB | E | 0.36 | 1 | — | — | — | — | — | — |
| CoMSIA | LB | H | 0.32 | 4 | — | — | — | — | — | — |
| CoMSIA | LB | D | 0.00 | 1 | — | — | — | — | — | — |
| CoMSIA | LB | A | 0.33 | 2 | — | — | — | — | — | — |
| CoMSIA | LB | E/S | 0.37 | 2 | — | — | — | — | — | — |
| CoMSIA | LB | 0.60E/0.40A | 0.41 | 2 | — | — | — | — | — | — |
| CoMSIA | LB | 0.73E/0.27D | 0.41 | 3 | — | — | — | — | — | — |
| CoMSIA | LB | 0.51E/.49H | 0.39 | 2 | — | — | — | — | — | — |
| CoMSIA | LB | 0.41E/0.27S/0.31A | 0.42 | 4 | 0.80 | 39.2 | 0.26 | 0.85 | 0.04 | 0.57 |
| CoMSIA | LB | E/A/D | 0.41 | 2 | — | — | — | — | — | — |
| CoMSIA | RG | S | 0.37 | 1 | — | — | — | — | — | — |
| CoMSIA | RG | E | 0.46 | 4 | — | — | — | — | — | — |
| CoMSIA | RG | H | 0.35 | 3 | — | — | — | — | — | — |
| CoMSIA | RG | D | 0.15 | 5 | — | — | — | — | — | — |
| CoMSIA | RG | A | 0.39 | 3 | — | — | — | — | — | — |
| CoMSIA | RG | 0.71E/0.29S | 0.46 | 5 | — | — | — | — | — | — |
| CoMSIA | RG | 0.66E/0.34A | 0.52 | 5 | 0.85 | 41.2 | 0.23 | 0.89 | 0.04 | 0.57 |
| CoMSIA | RG | E/D | 0.48 | 6 | — | — | — | — | — | — |
| CoMSIA | RG | 0.57E/0.43H | 0.48 | 5 | — | — | — | — | — | — |
| CoMSIA | RG | 0.54E/0.21S/0.25A | 0.51 | 5 | 0.86 | 45.4 | 0.22 | 0.9 | 0.03 | 0.64 |
| CoMSIA | RG | E/A/D | 0.47 | 6 | — | — | — | — | — | — |

Table 3: Statistics of different PLS analysis after region focusing. (GS: geometrical scheme; SE: standard error of estimate; n.: number of components; $F$: Fischer's $F$ value for test of significance; $r^2_{\mathrm{bs}}$: coefficient of determination after 100 bootstrapping runs; SD: standard deviation; Field contribution: (S) steric field, (E) electrostatic field, (H) hydrophobic field, (D) H-bond donor field, and (A) H-bond acceptor field.).

| Analysis | GS | Field | Grid spacing | $q^2$ | n. | $r^2$ | $F$ | SE | $r^2_{\mathrm{bs}}$ | SD | $r^2_{\mathrm{pred}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CoMFA | LB | S | 0.5 Å | 0.59 | 5 | 0.84 | 40.41 | 0.234 | 0.89 | 0.11 | 0.65 |
| CoMFA | LB | E | 0.5 Å | 0.13 | 2 | — | — | — | — | — | — |
| CoMFA | LB | 0.57S/0.43E | 0.5 Å | 0.56 | 4 | 0.84 | 49.62 | 0.235 | 076 | 0.13 | 0.68 |
| CoMFA | LB | S | 1.5 Å | 0.25 | 1 | 0.36 | 23.40 | 0.450 | — | — | — |
| CoMFA | LB | E | 1.5 Å | −0.03 | 1 | — | — | — | — | — | — |
| CoMFA | LB | 0.35S/0.65E | 1.5 Å | 0.38 | 2 | 0.51 | 21.20 | 0.40 | — | — | — |
| CoMFA | RG | S | 0.5 Å | 0.41 | 4 | 0.71 | 23.47 | 0.315 | — | — | — |
| CoMFA | RG | E | 0.5 Å | 0.42 | 7 | 0.87 | 33.11 | 0.220 | — | — | — |
| CoMFA | RG | 0.47S/0.53E | 0.5 Å | 0.55 | 6 | 0.94 | 97.78 | 0.145 | 0.96 | 0.02 | 0.67 |
| CoMFA | RG | S | 1.5 Å | 0.44 | 4 | 0.65 | 17.93 | 0.345 | — | — | — |
| CoMFA | RG | E | 1.5 Å | 0.11 | 4 | 0.26 | 3.32 | 0.505 | — | — | — |
| CoMFA | RG | 0.45S/0.50E | 1.5 Å | 0.29 | 3 | 0.60 | 18.99 | 0.370 | — | — | — |

Table 4: Experimental and predicted activities with their residuals by CoMFA and CoMSIA analyses of the training set.

| n. | CoMFA | | | CoMSIA | |
|---|---|---|---|---|---|
| | Experimental | Predicted | | Predicted | |
| | pIC$_{50}$ | pIC$_{50}$ | Residual | pIC$_{50}$ | Residual |
| 1 | 5.706 | 5.553 | 0.153 | 5.646 | 0.060 |
| 2 | 6.286 | 6.296 | −0.010 | 6.172 | 0.114 |
| 3 | 5.870 | 5.948 | −0.078 | 5.781 | 0.089 |
| 4 | 5.480 | 5.888 | −0.408 | 5.814 | −0.334 |
| 6 | 5.979 | 5.365 | 0.614 | 5.421 | 0.558 |
| 9 | 5.202 | 5.275 | −0.073 | 5.074 | 0.128 |
| 10 | 5.963 | 6.200 | −0.237 | 6.199 | −0.236 |
| 11 | 6.218 | 6.191 | 0.027 | 6.135 | 0.083 |
| 12 | 6.650 | 6.276 | 0.374 | 6.330 | 0.320 |
| 13 | 6.292 | 6.065 | 0.227 | 5.849 | 0.443 |
| 14 | 5.910 | 5.870 | 0.040 | 5.842 | 0.068 |
| 15 | 6.465 | 6.487 | −0.022 | 6.363 | 0.102 |
| 16 | 5.674 | 5.927 | −0.253 | 5.787 | −0.113 |
| 18 | 5.484 | 5.473 | 0.011 | 5.589 | −0.105 |
| 19 | 5.238 | 5.102 | 0.136 | 5.579 | −0.341 |
| 20 | 5.550 | 5.615 | −0.065 | 5.625 | −0.075 |
| 21 | 6.047 | 6.015 | 0.032 | 6.226 | −0.179 |
| 22 | 5.349 | 5.544 | −0.195 | 5.754 | −0.405 |
| 24 | 6.668 | 6.571 | 0.097 | 6.474 | 0.194 |
| 25 | 5.870 | 6.034 | −0.164 | 5.832 | 0.038 |
| 26 | 5.777 | 5.929 | −0.152 | 6.416 | −0.639 |
| 27 | 6.606 | 6.586 | 0.02 | 6.498 | 0.108 |
| 28 | 5.979 | 5.981 | −0.002 | 5.676 | 0.303 |
| 29 | 5.455 | 5.471 | −0.016 | 5.562 | −0.107 |
| 30 | 5.421 | 5.473 | −0.052 | 5.604 | −0.183 |
| 32 | 6.656 | 6.366 | 0.290 | 6.297 | 0.359 |
| 33 | 6.111 | 6.366 | −0.255 | 6.297 | −0.186 |
| 35 | 5.740 | 5.706 | 0.034 | 6.068 | −0.328 |
| 37 | 5.886 | 5.934 | −0.048 | 5.864 | 0.022 |
| 38 | 5.857 | 5.828 | 0.029 | 5.763 | 0.094 |
| 39 | 6.924 | 6.826 | 0.098 | 6.785 | 0.139 |
| 40 | 6.642 | 6.826 | −0.184 | 6.785 | −0.143 |
| 41 | 6.939 | 7.009 | −0.070 | 6.951 | −0.012 |
| 42 | 7.092 | 7.032 | 0.060 | 6.916 | 0.176 |
| 43 | 6.987 | 7.012 | −0.025 | 7.008 | −0.021 |
| 46 | 7.051 | 6.950 | 0.101 | 7.130 | −0.079 |
| 47 | 7.222 | 7.325 | −0.103 | 7.126 | 0.096 |
| 48 | 6.780 | 6.763 | 0.017 | 6.881 | −0.101 |
| 50 | 6.830 | 6.776 | 0.054 | 6.842 | −0.012 |
| 51 | 6.658 | 6.648 | 0.010 | 6.806 | −0.148 |
| 52 | 6.577 | 6.547 | 0.030 | 6.429 | 0.148 |
| 53 | 6.279 | 6.328 | −0.049 | 6.298 | −0.019 |
| 54 | 6.256 | 6.328 | −0.072 | 6.298 | −0.042 |

Table 5: Experimental and predicted activities with their residuals by CoMFA and CoMSIA analyses of the test set.

| n. | CoMFA<br>Experimental<br>pIC$_{50}$ | Predicted<br>pIC$_{50}$ | Residual | CoMSIA<br>Predicted<br>pIC$_{50}$ | Residual |
|---|---|---|---|---|---|
| 5 | 5.456 | 5.404 | 0.052 | 5.536 | −0.080 |
| 7 | 5.644 | 5.779 | −0.135 | 5.811 | −0.167 |
| 8 | 5.955 | 6.015 | −0.060 | 5.787 | 0.168 |
| 17 | 6.009 | 5.966 | 0.043 | 5.819 | 0.190 |
| 23 | 6.483 | 6.274 | 0.209 | 6.304 | 0.179 |
| 31 | 5.775 | 5.703 | 0.072 | 5.728 | 0.047 |
| 34 | 6.979 | 6.269 | 0.710 | 6.205 | 0.774 |
| 36 | 5.469 | 5.696 | −0.227 | 5.924 | −0.455 |
| 44 | 7.041 | 7.080 | −0.039 | 7.227 | −0.186 |
| 45 | 6.936 | 6.950 | −0.014 | 7.130 | −0.194 |
| 49 | 6.625 | 6.902 | −0.277 | 6.834 | −0.209 |

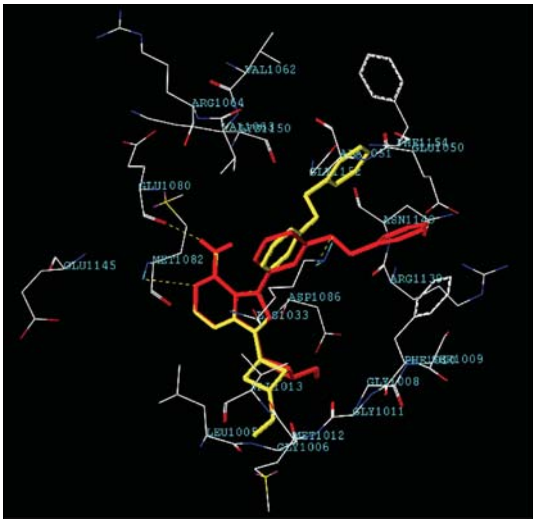


Figure 1: Comparison of minimum energy (yellow) and docking based (red) conformers.

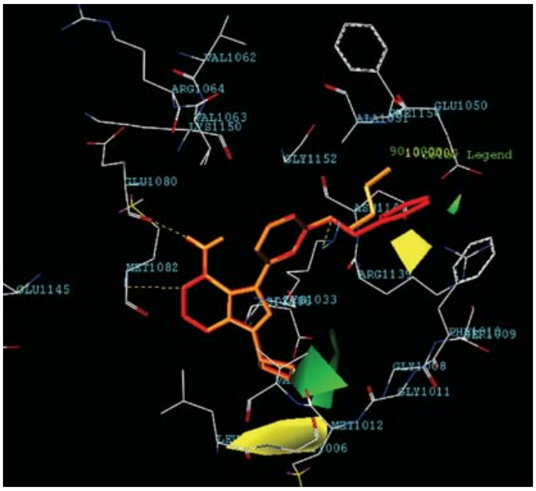


Figure 3: CoMFA steric maps with the most (red) and least (orange) active compound within the active site.

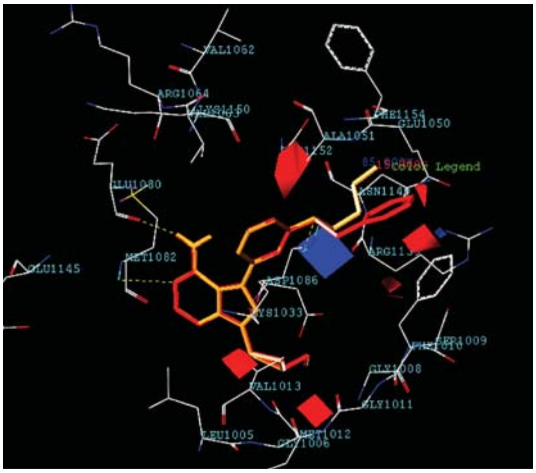


Figure 2: CoMFA electrostatic maps with the most (red) and least (orange) active compound within the active site.

red- and blue-colored contours while steric interactions are represented by green and yellow colored contours. In electrostatic field, blue color contour represents region where electropositive group enhances the activity, whereas red-color region likes electron-rich groups to increase the biological activity. In case of steric interactions, the green region demands bulky substituents to enhance the activity, while in yellow contours, bulky substituents decrease the activity.

The most potent compound-47 (red color) and least-active compound-9 (orange color) of the series with CoMFA contour maps have been superimposed in the active site of the receptor protein. Figure 2 shows that red polyhedrons locate the region where electron-rich group will enhance the inhibitory activity, and vice versa for blue polyhedron. Therefore, the phenyl ring in compound-47 might be responsible for its higher activity than methoxy group of compound-9 because it might have the $\pi$-$\pi$ interactions with the phenyl ring of phenyl alanine (Phe1010) amino acid. The red contour around 1–3 carbon of cyclobutane also demands the electron-rich group for higher potency. Compound-47 has amino group at C-3 position which might be responsible for its higher activity than least-active compound-9. It is also clear in most of compounds from the dataset that electron-rich group at this position have higher activity than compound-9. In Figure 3, green polyhedron locates the region where bulky substitutent would increase the inhibitory activity and yellow polyhedron where the steric bulk is not required for high potency of the compounds. The small green contour near the phenyl ring of compound-47 explains its higher activity than compound-9. Similarly, the green contour around 2 and 3 carbon of cyclobutane requires the bulky substitutent to be highly active. Thus the bulky substitutent at this position in dataset favors the higher inhibitory activity of the compounds than compound-9. Yellow polyhedron below the plane of phenyl ring and cyclopropane requires the small group to be more active.

### 3.6. *CoMSIA contour maps*

The CoMSIA contour maps were also developed on the models based on the geometrical scheme 2. Figures 4, 5, and 6 show the steric electrostatic and H-bond acceptor contour maps superimposed in the active site of the IGF-1R. In CoMSIA method, steric and electrostatic contours maps have the same meaning as that of CoMFA contour maps whereas H-bond acceptor contours are represented by magenta and red colors. Magenta favors H-bond acceptor group while red disfavors. The steric and electrostatic maps are more or less similar to the corresponding CoMFA models (Figures 2 and 3, resp.) except that there is a small green contour near phenyl ring of compound-47 in CoMFA model. In Figure 6, the magenta contour around C-2 and C-3 position of cyclobutane favors the H-bond accepting group to enhance the inhibitory activity of the molecules. Thus the H-bond accepting substituent at C-4 position might enhance inhibitory activity of the compounds through H-bonding with Glycine (Gly1008) or Valine (Val1013).

## 4. CONCLUSION

A comparative CoMFA and CoMSIA models were developed for the series of potent IGF-1R inhibitors. Ligand-based and receptor-guided protocols were applied to develop the models. Receptor-guided alignment gave models with better statistics than the ones from the ligand-based approach, presumably because the alignment using receptor information is more realistic. Moreover, the interpretation of receptor-guided models are directly associated with the receptor information. That is, in general, the superposition of a CoMFA or CoMSIA contour map inside the receptor shows reasonable correspondence between the contour map property and the physical property of surrounding active site region. This provides more detailed understanding about the interaction between the series of inhibitors and IGF-1R. The information drawn here can be used to design new inhibitors of IGF-1R.

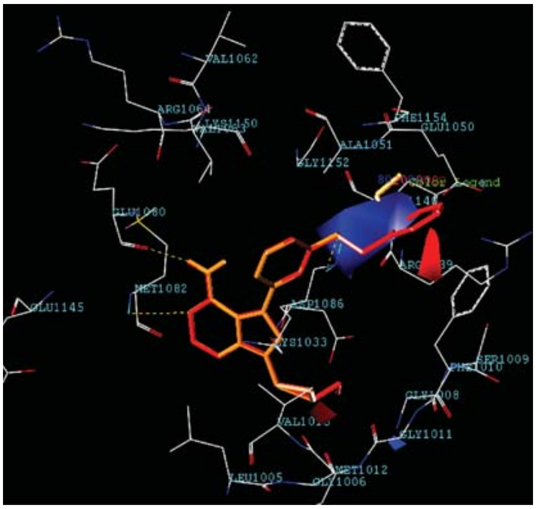


FIGURE 4: CoMSIA electrostatic maps with the most (red) and least (orange) active compound within the active site.

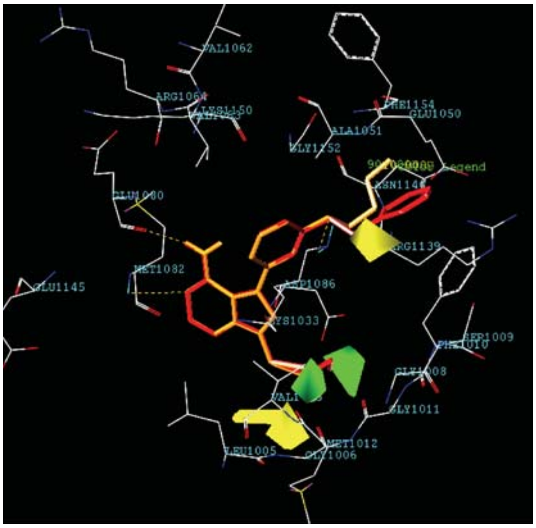


FIGURE 5: CoMSIA steric maps with the most (red) and least (orange) active compound within the active site.

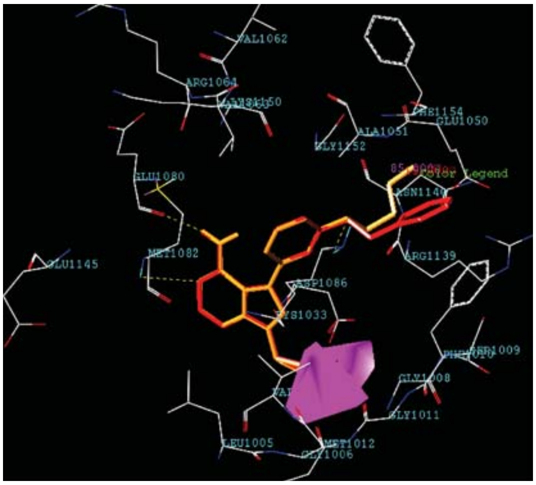


FIGURE 6: CoMSIA H-bond acceptor map with the most (red) and least (orange) active compound within the active site.

## ACKNOWLEDGMENT

The authors would like to thank Jung Soo Oh for the valuable support.

## REFERENCES

[1] V. M. Macaulay, "Insulin-like growth-factors and cancer," *British Journal of Cancer*, vol. 65, no. 3, pp. 311–320, 1992.

[2] C. L. Arteaga and C. K. Osborne, "Growth inhibition of human breast cancer cells in vitro with an antibody against the type-I somatomedin receptor," *Cancer Research*, vol. 49, no. 22, pp. 6237–6241, 1989.

[3] H. Werner and D. Le Roith, "New concepts in regulation and function of the insulin-like growth factors: implications for understanding normal growth and neoplasia," *Cellular and Molecular Life Sciences*, vol. 57, no. 6, pp. 932–942, 2000.

[4] S. E. Hankinson, W. C. Willett, G. A. Colditz, et al., "Circulating concentrations of insulin-like growth factor-I and risk of breast cancer," *Lancet*, vol. 351, no. 9113, pp. 1393–1396, 1998.

[5] C. García-Echeverría, M. A. Pearson, A. Marti, et al., "In vivo antitumor activity of NVP-AEW541—a novel, potent, and selective inhibitor of the IGF-IR kinase," *Cancer Cell*, vol. 5, no. 3, pp. 231–239, 2004.

[6] R. Kuttan, P. Bhanumathy, K. Nirmala, and M. C. George, "Potential anticancer activity of turmeric (Curcuma longa)," *Cancer Letters*, vol. 29, no. 2, pp. 197–202, 1985.

[7] P. Haluska, J. M. Carboni, D. A. Loegering, et al., "In vitro and in vivo antitumor effects of the dual insulin-like growth factor-I/insulin receptor inhibitor, BMS-554417," *Cancer Research*, vol. 66, no. 1, pp. 362–371, 2006.

[8] F. A. Pasha, H. W. Chung, S. J. Cho, and S. B. Kang, "3D-quantitative structure activity analysis and quantum chemical analysis of pyrido-di-indoles," *International Journal of Quantum Chemistry*, vol. 108, no. 2, pp. 391–400, 2008.

[9] F. A. Pasha, K. Dal Nam, and S. J. Cho, "CoMFA based quantitative structure toxicity relationship of azo dyes," *Molecular & Cellular Toxicology*, vol. 3, no. 2, pp. 145–149, 2007.

[10] F. A. Pasha, M. M. Neaz, S. J. Cho, and S. B. Kang, "Quantitative structure activity relationship (QSAR) study of estrogen derivatives based on descriptors of energy and softness," *Chemical Biology & Drug Design*, vol. 70, no. 6, pp. 520–529, 2007.

[11] F. A. Pasha, H. K. Srivastava, A. Srivastava, and P. P. Singh, "QSTR study of small organic molecules against Tetrahymena pyriformis," *QSAR and Combinatorial Science*, vol. 26, no. 1, pp. 69–84, 2007.

[12] G. Klebe, U. Abraham, and T. Mietzner, "Molecular similarity indexes in a comparative-analysis (Comsia) of drug molecules to correlate and predict their biological-activity," *Journal of Medicinal Chemistry*, vol. 37, no. 24, pp. 4130–4146, 1994.

[13] G. Klebe, U. Abraham, and T. Mietzner, "Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity," *Journal of Medicinal Chemistry*, vol. 37, no. 24, pp. 4130–4146, 1994.

[14] R. D. Cramer, D. E. Patterson, and J. D. Bunce, "Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins," *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988.

[15] M. J. Mulvihill, Q. S. Ji, D. Werner, et al., "1,3-disubstituted-imidazo[1,5-a]pyrazines as insulin-like growth-factor-I receptor (IGF-IR) inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 4, pp. 1091–1097, 2007.

[16] K. H. Kim, G. Greco, and E. Novellino, "A critical review of recent CoMFA applications," *Perspectives in Drug Discovery and Design*, vol. 12–14, pp. 257–315, 1998.

[17] C. Kunick, K. Lauenroth, K. Wieking, et al., "Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones," *Journal of Medicinal Chemistry*, vol. 47, no. 1, pp. 22–36, 2004.

[18] D. C. Juvale, V. V. Kulkarni, H. S. Deokar, N. K. Wagh, S. B. Padhye, and V. M. Kulkarni, "3D-QSAR of histone deacetylase inhibitors: hydroxamate analogues," *Organic and Biomolecular Chemistry*, vol. 4, no. 15, pp. 2858–2868, 2006.

[19] W. Sippl, "Receptor-based 3D QSAR analysis of estrogen receptor ligands—merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 6, pp. 559–572, 2000.

[20] A. Pautsch, A. Zoephel, H. Ahorn, W. Spevak, R. Hauptmann, and H. Nar, "Crystal structure of bisphosphorylated IGF-1 receptor kinase: insight into domain movements upon kinase activation," *Structure*, vol. 9, no. 10, pp. 955–965, 2001.

[21] M. Clark, R. D. Cramer, and N. Vanopdenbosch, "Validation of the general-purpose tripos 5.2 force-field," *Journal of Computational Chemistry*, vol. 10, no. 8, pp. 982–1012, 1989.

[22] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, "The collinearity problem in linear-regression—the partial least-squares (Pls) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[23] A. Golbraikh and A. Tropsha, "Beware of q2!," *Journal of Molecular Graphics and Modelling*, vol. 20, no. 4, pp. 269–276, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 860270, 10 pages
doi:10.1155/2008/860270

*Research Article*

# An Algorithm for Finding Functional Modules and Protein Complexes in Protein-Protein Interaction Networks

**Guangyu Cui,[1] Yu Chen,[1] De-Shuang Huang,[2] and Kyungsook Han[1]**

[1] *School of Computer Science and Engineering, Inha University, Incheon 402-751, South Korea*
[2] *Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, China*

Correspondence should be addressed to Kyungsook Han, khan@inha.ac.kr

Received 1 September 2007; Revised 12 November 2007; Accepted 24 December 2007

Recommended by Daniel Howard

Biological processes are often performed by a group of proteins rather than by individual proteins, and proteins in a same biological group form a densely connected subgraph in a protein-protein interaction network. Therefore, finding a densely connected subgraph provides useful information to predict the function or protein complex of uncharacterized proteins in the highly connected subgraph. We have developed an efficient algorithm and program for finding cliques and near-cliques in a protein-protein interaction network. Analysis of the interaction network of yeast proteins using the algorithm demonstrates that 59% of the near-cliques identified by our algorithm have at least one function shared by all the proteins within a near-clique, and that 56% of the near-cliques show a good agreement with the experimentally determined protein complexes catalogued in MIPS.

## 1. INTRODUCTION

Proteins in a highly connected subgraph of a protein interaction network usually share a common function [1]. Therefore, a highly connected subgraph such as clique and near-clique in a protein interaction network can be used to predict the function of uncharacterized proteins in the highly connected subgraph. Finding a clique with a maximum size in a graph is an NP-hard problem [2]. There are several heuristic algorithms for the maximum clique problem [2, 3], but most of them focus on finding a complete subgraph (i.e., clique) and cannot be used to find near-cliques.

Several topological analysis methods have been developed for identifying biologically meaningful groups from protein interaction networks or for assessing the reliability of protein interactions. A recent program called CFinder [4, 5] finds overlapping cliques in protein interaction networks. It allows a protein to belong to more than one clique, but cannot find near-cliques. Our study shows that the near-cliques can reveal higher functional coherence than the overlapping cliques.

The primary focus of this study is to find functional groups by identifying cliques and near-cliques in protein interaction networks. This study attempts to answer two questions as follows. "Can we efficiently find all cliques and near-cliques?" and "does a dense subgraph such as clique and near-clique indeed represent a functional module or protein complex?" This study demonstrates that the answers to both questions are "yes." This paper presents an algorithm for finding near-cliques and its application to the interaction network of yeast proteins.

## 2. ALGORITHMS FOR FINDING NEAR-CLIQUES

A clique is a complete graph $G = (N, E)$ in which every node is connected to every other node in the graph. In our previous work, we developed a heuristic algorithm and implemented the algorithm in a program called InterViewer [6], which identifies all edge-disjoint cliques (i.e., cliques that do not share an edge).

Our experience with protein interaction networks suggests that a near-clique as well as a clique often represents a biologically meaningful unit such as functional module or protein complex. A near-clique is almost a clique but is not a clique due to a few missing edges. We consider near-cliques of the following *basic* types, which are biologically meaningful clusters (see Figure 1).

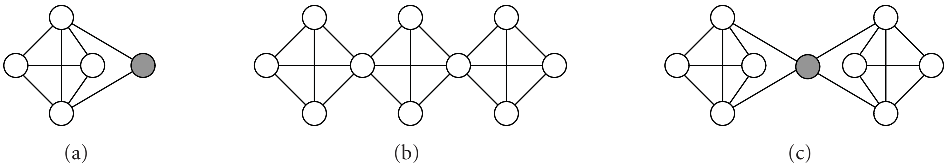


FIGURE 1: Near-cliques of types A, B, and C. Proteins outside a clique are represented as shaded nodes.

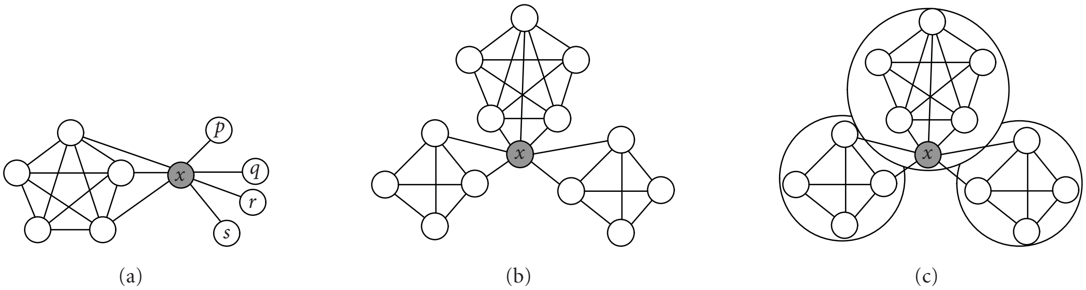


FIGURE 2: (a) After removing nodes $p$, $q$, $r$, and $s$ and their edges, node $x$ forms a near-clique of type A with the remaining nodes. (b) This graph becomes a near-clique $G$ of type C since indegree$(x, G) \geq 0.5|G|$. (c) A big near-clique is too big (e.g., near-clique with more than 50 nodes) and is split into smaller near-cliques (in this example, 3 small near-cliques).

### *Type A*

When a protein outside a clique interacts with two or more proteins in the clique, the protein and the clique forms a near-clique.

### *Type B*

When a clique shares a protein with other cliques, the cliques form a near-clique.

### *Type C*

When two or more cliques interact with a common protein outside them and the protein has at least two interactions with each clique, the cliques and the protein form a near-clique.

The near-cliques of types A and C can be refined using the indegree and outdegree of a node (there is no change to the near-clique of type B). For a node $x$ in subgraph $G' \subset G$, indegree$(x, G')$ is the number of the edges connecting node $x$ to other nodes in $G'$, and outdegree$(x, G')$ is the number of edges connecting node $x$ to other nodes that are in $G$ but not in $G'$. We use the definition of a community in a strong sense [7] to find more near-cliques in a graph.

*Definition 1.* A subgraph $G'$ is a community in a strong sense if indegree$(x, G') >$ outdegree$(x, G')$ for every $x$ in $G'$.

The original definition of a strong community misses many near-cliques due to a single node in the communities. For example, in Figure 2(a), node $x$ cannot belong to a near-clique since indegree$(x, G') = 3 <$ outdegree$(x, G') = 4$. Likewise, node $x$ in Figure 2(b) cannot belong to a near-clique because indegree$(x, G') <$ outdegree$(x, G')$. Thus, nodes with only one edge connected to them and their edges are removed from the graph when we search near-cliques in the graph. In the graph of Figure 2(a), nodes $p$, $q$, $r$, and $s$ and their edges are removed. After removing them, node $x$ and the existing clique form a near-clique of type A. A cluster that satisfies indegree$(x, G') \geq 0.5|G'|$ for every $x$ in $G'$, where $|G'|$ is the number of nodes in $G'$, forms a near-clique, too.The example shown in Figure 2(b) becomes a near-clique since it satisfies indegree$(x, G') \geq 0.5|G'|$ even if it does not satisfy indegree$(x, G') <$ outdegree$(x, G')$.

Therefore, a near-clique $G$ of basic types A and C should satisfy at least one of the following conditions.

(1) indegree$(x, G) \geq$ outdegree$(x, G)$ for every $x$ in $G$.

(2) indegree$(x, G) \geq 0.5|G|$.

After finding all edge-disjoint cliques first, we identify near-cliques as follows. More detailed description of finding near-cliques are outlined in Algorithms 1 and 2. In the algorithms, cIdx represents the index of a clique.

(1) Assign every node of a clique the index of the clique containing the node.

(2) When a node of a clique has already an assigned clique index, assign the index to all nodes of the clique, and merge two cliques into a near-clique of type B.

(3) When a node $x$ outside a clique forms a basic near-clique $G$ of type A due to the interactions with two or more proteins in the clique, and either indegree$(x, G) \geq$ outdegree$(x, G)$ or indegree $i(x, G) \geq 0.5|G|$ is true, assign the index of the clique to the node.

(4) When two or more cliques form a near-clique $G$ due to two or more interactions with a common protein outside the cliques, and either indegree$(x, G) \geq$ outdegree$(x, G)$ or indegree $i(x, G) \geq 0.5|G|$ is true, merge the cliques and the protein into a near-clique of type C. A near-clique is formed by selecting nodes with the same clique index ($cIdx$) as those nodes with $cIdx > 0$.

```
for all node $N \in G$ do
    $N.cIdx = 0$                                   {initialize cIdx of all nodes to 0}
end for
$curCIdx = 1$                                      {set the current clique index to 1}
for all node $N \in G$ do
    if (isClique($N$)) then {if the node $N$ belongs to a clique}
        for all edge $E \in N$ do
            if ($E$.target.$cIdx > 0$) then {if the $cIdx$ of the node connected to $N$ is positive}
                for all $tmpN \in G$ do {for all nodes in $G$}
                    if ($tmpN.cIdx = E$.target.$cIdx$) then
                        $tmpN.cIdx = curCIdx$      {assign $curCIdx$ to $tmpN$ as its cIdx}
                    end if
                end for
            else
                $E$.target.$cIdx = curCIdx$
            end if
        end for
        $N.cIdx = curCIdx$                         {set $cIdx$ of $N$ to $curCIdx$}
        $curCIdx + +$                              {increase $curCIdx$ by one}
    end if
end for
```

Algorithm 1: AssignNearCliqueIdx.

```
for all node $N \in G$ do
    if ($N.cIdx = 0$) then                {find node outside the clique}
        $qCliqueCnts = \varnothing$ {$qCliqueCnts$ the number of edges, which the node N connected with different near-cliques}
        for all edge $E \in N$ do
            if ($E$.target.$cIdx > 0$) then
                $qCliqueCnts[E$.target.$cIdx] + +$
            end if
        end for
        $qC$ value $= 0$                                              {initialize $cIdx$ of node N}
        for all ($c \in qCliqueCnts$) do
            if (($c > 1$) and indegree($x, G'$) $\geq$ outdegree($x, G'$)) or (($c > 1$) and indegree($x, G'$) $\geq 0.5 * |G'|$) then {a node
            outside a clique interacts with multiple nodes in the clique, and either indegree($x, G'$) $\geq$ outdegree($x, G'$) or
            indegree($x, G'$) $\geq 0.5 * |G'|$) is true}
                if ($qC$ value $> 0$) then
                    for all $tmpN \in G$ do
                        if ($tmpN.cIdx = qC$ value) then
                            $tmpN.cIdx = qC$ value                           {near-clique of type C}
                        end if
                    end for
                else
                    $qC$ value $= c$                                      {near-clique of type A}
                end if
            end if
        end for
        $N.cIdx = qC$ value                          {assign $qC$ value to node $N$ as its $cIdx$}
    end if
end for
```

Algorithm 2: ExtendNearClique.

Since the most relevant processes form a group of proteins of moderate size in biological networks [8], we obtain near-cliques smaller than the maximum size specified by a user. That is, when a near-clique bigger than the maximum size is found (e.g., near-clique with more than 50 nodes), it is split into smaller near-cliques (3 near-cliques in Figure 2(c)). The way we split a big near-clique is as follows. When our program finds a big near-clique with the minimum clique size set to $k$, we rerun the program on the big near-clique with the minimum clique size set to $k + 1$ to find a new

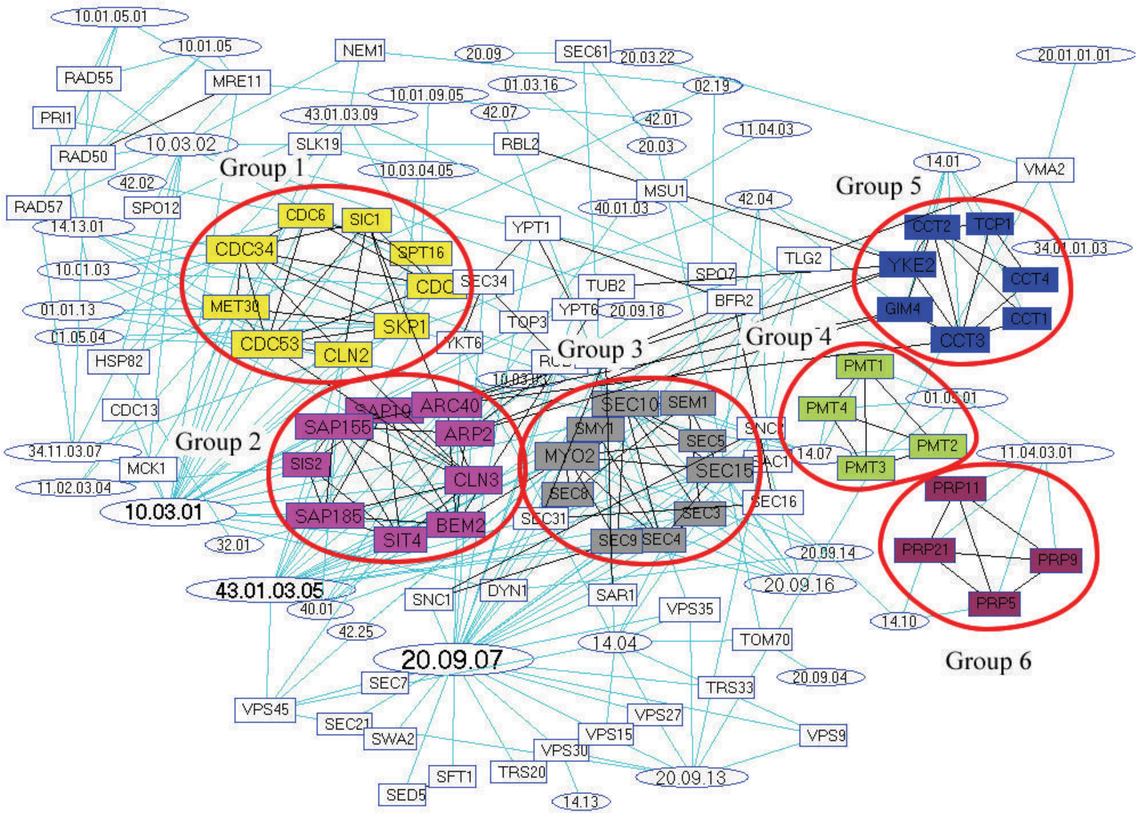


FIGURE 3: Six near-cliques found in yeast protein interaction networks. Proteins in each near-clique share at least one function with other proteins within the near-clique.

clique and a near-clique with the clique. After removing the new near-clique from the original, big near-clique, we run the program again with the minimum clique size set to $k$. The big near-clique shown in Figure 2(c) is split into 3 small near-cliques with at least 4 proteins each.

## 3. RESULTS AND COMPARISON WITH EXPERIMENTAL DATA

We tested the algorithms on the data with 8,397 interactions between 4,380 yeast proteins, which is the combined data of Ito et al. [9], Uetz et al. [10], and MIPS (http://mips.gsf.de) with redundant data removed. To every protein in the near-cliques, we assigned the functional categories of the Functional Catalog (FunCat) version 2.0 [11], which includes 97 functional categories. There are six levels of hierarchy in the FunCat structure.

In the data with 8,397 interactions between 4,380 yeast proteins, we found 100 near-cliques with the minimum size of a clique set to 3 and the maximum size of a near-clique set to 40. Only one near-clique contains more than 40 proteins, and so it was split into 17 small near-cliques, resulting in total 116 near-cliques. Figure 3 shows an example of the network of yeast protein interactions with 6 near-cliques. Proteins in each near-clique share at least one function with other proteins within the near-clique.

As shown in Table 1, 68 (59%) out of the 116 near-cliques have at least one function shared by all the proteins in the near-cliques (100% sharing), and 39 near-cliques have a function shared by more than 50% of the proteins in the near-cliques, supporting data are available at http://wilab.inha.ac.kr/ppi/homepage.mht. Only 9 near-cliques have no function shared by >50% of the proteins in the near-cliques. As shown in Figure 4, the functional coherence of each near-clique is high. The functional coherence was computed by the ratios of the number of proteins having a specific functional category to the group size (i.e., the number of proteins in the group).

Interestingly, most near-cliques found by our algorithm belong to multifunctional categories. For example, two functional categories are common to all the proteins in a near-clique of Figure 5. As shown in Table 2, the near-clique identified as group 93 by our program is involved in both stress response (functional category 32.01) and biosynthesis of vitamins, cofactors, and prosthetic groups (functional category 01.07.01).

Near-cliques may correspond to protein complexes in addition to functional modules. So, we compared the near-cliques identified by our algorithms with known yeast protein complexes, which are cataloged in the MIPS Saccharomyces cerevisiae genome database (http://mips.gsf.de/genre/proj/yeast). For each near-clique, we found a best-matching protein complex by minimizing

Table 1: Functional groups identified from the yeast protein interaction data. 68 modules have at least one function shared by all the proteins in the groups (100% sharing), and 39 groups have a function shared by more than 50% of the proteins in the groups. Only 9 groups have no function shared by >50% of the proteins in the group. This table shows only one function with the highest functional coherence in each group. All the functions shared by more than 50% of the proteins in each group are available at http://wilab.inha.ac.kr/ppi/homepage.mht.

| Group ID | Proteins in the group | Common function (proportion of proteins with the function) | Group ID | Proteins in the group | Common function (proportion of proteins with the function) | Group ID | Proteins in the group | Common function (proportion of proteins with the function) |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 11.02.02 (100%) | 40 | 3 | 10.01.03.03 (100%) | 79 | 3 | 42.01 (100%) |
| 2 | 5 | 14.01 (100%) | 41 | 3 | 20.09.13 (100%) | 80 | 3 | 30.01.05.01 (66.7%) |
| 3 | 4 | 42.10 (100%) | 42 | 5 | 42.01 (80%) | 81 | 3 | 42.04.03 (100%) |
| 4 | 12 | 42.10.05 (75%) | 43 | 5 | 40.01 (80%) | 82 | 4 | 42.25 (100%) |
| 5 | 4 | 01.03.16.01 (75%) | 44 | 3 | 14.10 (100%) | 83 | 3 | none |
| 6 | 6 | 11.02.03.04 (100%) | 45 | 8 | 12.04.01 (87.5%) | 84 | 3 | 01.06.01.07.11 (100%) |
| 7 | 4 | 14.07.02.01 (100%) | 46 | 4 | 14.13.01.01 (100%) | 85 | 3 | 32.01.07 (66.7%) |
| 8 | 6 | 10.03.01 (100%) | 47 | 4 | 10.01.05.01 (100%) | 86 | 4 | 34.11.03.07 (100%) |
| 9 | 22 | 11.04.01 (63.6%) | 48 | 3 | 10.01.05.01 (100%) | 87 | 3 | 20.09.07 (100%) |
| 10 | 21 | 20.09 (66.7%) | 49 | 6 | 20.09.04 (100%) | 88 | 3 | 02.19 (100%) |
| 11 | 3 | none | 50 | 4 | 32.01 (100%) | 89 | 3 | 16.19.03 (100%) |
| 12 | 8 | 11.04.03.05 (100%) | 51 | 3 | 10.01.03 (100%) | 90 | 4 | 2.07 (75%) |
| 13 | 11 | 10.03.01 (63.6%) | 52 | 3 | 12.04.03 (66.7%) | 91 | 3 | 16.03.01 (100%) |
| 14 | 7 | 10.03.01 (76.5%) | 53 | 13 | 20.09.07.03 (61.5%) | 92 | 3 | 10.01.05.01 (100%) |
| 15 | 4 | 1.03 (50%) | 54 | 8 | 11.02.03.01 (100%) | 93 | 7 | 32.01 (100%) |
| 16 | 5 | 01.05.01.03.02.02 (100%) | 55 | 4 | 20.09.07.03 (100%) | 94 | 3 | 40.20 (66.7%) |
| 17 | 3 | 16.03.01 (100%) | 56 | 5 | none | 95 | 3 | 34.01.01.03 (100%) |
| 18 | 4 | 11.04.02 (100%) | 57 | 5 | 20.09.01 (100%) | 96 | 4 | 43.01.03.05 (100%) |
| 19 | 5 | 40.01 (80%) | 58 | 5 | 20.09.18 (80%) | 97 | 3 | 14.04 (100%) |
| 20 | 3 | 18.02.01 (60%) | 59 | 5 | 01.04.01 (80%) | 98 | 4 | 20.09.13 (100%) |
| 21 | 23 | 43.01.03.05 (82.6%) | 60 | 3 | 43.01.03.05 (100%) | 99 | 3 | 16.03.01 (66.7%) |
| 22 | 4 | 32.01 (50%) | 61 | 5 | 11.04.01 (100%) | 100 | 12 | 43.01.03.05 (91.7%) |
| 23 | 4 | 11.02.03.04.01 (100%) | 62 | 5 | 20.09.10 (100%) | 101 | 6 | 10.03.01.01.03 (100%) |
| 24 | 4 | 14.13.01.01 (100%) | 63 | 3 | 43.01.03.09 (66.7%) | 102 | 9 | 16.01 (88.9%) |
| 25 | 36 | none | 64 | 3 | 11.02.03.04 (100%) | 103 | 7 | 43.01.03.05 (100%) |

Table 1: Continued.

| Group ID | Proteins in the group | Common function (proportion of proteins with the function) | Group ID | Proteins in the group | Common function (proportion of proteins with the function) | Group ID | Proteins in the group | Common function (proportion of proteins with the function) |
|---|---|---|---|---|---|---|---|---|
| 26 | 4 | 20.09.07.03 (100%) | 65 | 3 | 34.11.03.13 (100%) | 104 | 5 | 43.01.03.05 (80%) |
| 27 | 10 | 42.04 (50%) | 66 | 5 | 16.19.03 (80%) | 105 | 11 | 20.09.07.03 (100%) |
| 28 | 11 | 14.13.01.01 (100%) | 67 | 4 | 11.04 (100%) | 106 | 7 | 10.03.04.03 (85.7%) |
| 29 | 6 | 43.01.03.05 (83.3%) | 68 | 3 | 10.01.09.05 (66.7%) | 107 | 6 | 40.01 (50%) |
| 30 | 8 | 12.04 (100%) | 69 | 4 | 01.04.01 (100%) | 108 | 9 | 10.03.01.01 (88.9%) |
| 31 | 4 | 10.03.01 (100%) | 70 | 3 | 11.06.01 (100%) | 109 | 3 | 20.09.07.27 (100%) |
| 32 | 4 | 12.04.02 (75%) | 71 | 6 | none | 110 | 3 | 20.09.14 (100%) |
| 33 | 3 | 34.01.01.01 (100%) | 72 | 3 | 20.09.13 (100%) | 111 | 12 | 43.01.03.05 (100%) |
| 34 | 5 | none | 73 | 3 | 11.02.03.04 (100%) | 112 | 5 | 10.03.04.05 (100%) |
| 35 | 3 | 11.04.01 (66.7%) | 74 | 3 | 42.10.03 (100%) | 113 | 5 | 10.03.04.05 (100%) |
| 36 | 31 | 11.02.03.04 (80.6%) | 75 | 3 | none | 114 | 5 | 42.10.03 (80%) |
| 37 | 4 | 16.03.01 (100%) | 76 | 7 | 20.09.04 (100%) | 115 | 3 | none |
| 38 | 6 | 43.01.03.05 (100%) | 77 | 3 | none | 116 | 5 | 43.01.03.05 (80%) |
| 39 | 7 | 14.04 (57.1%) | 78 | 3 | 10.03.02 (100%) | — | — | — |

Table 2: Functional annotation of group 93 shown in Figure 5. The code represents functional category.

| Group | Protein | Protein functional categories | |
|---|---|---|---|
| Group 93 | YFL059w | 32.01 | 01.07.01 |
| | YMR096w | 32.01 | 01.07.01 |
| | YMR322c | 32.01.07 | 01.07 |
| | YNL334c | 32.01 | 01.07.01 |
| | YMR095c | 32.01 | 01.07.01 |
| | YNL333w | 32.01 | 01.07.01 |
| | YFL060c | 32.01 | 01.07.01 |

the probability of a random overlap between the two, using the following equation [4, 5]:

$$P_{\text{overlap}} = \frac{\binom{n_2}{k}\binom{N-n_2}{n_1-k}}{\binom{N}{n_1}}, \quad (1)$$

where $n1$, $n2$ are the sizes of a known protein complex and a computed module, $k$ is the number of their common proteins, and $N$ is the size of the network.

As shown in Table 3, 65 near-cliques (56% of the total 116 near-cliques) identified by our algorithm show a good agreement ($\ln(P_{\text{overlap}}) < -14$) with the protein complexes cataloged in MIPS.

To compare the functional coherence of the groups found by our program with that of cliques found by CFinder, we tested both programs on the same dataset. 75.9% of the groups identified by our program have at least two functional categories shared by all the proteins in the groups, whereas 63.1% of the groups identified by CFinder have at least two functional categories shared by all the proteins in the groups (Table 4). This result indicates that our program finds groups with stronger functional coherence than CFinder.

Table 5 shows the actual running times of our program and CFinder on three datasets of yeast protein interactions. Our program is faster than CFinder on all datasets, and the difference in speed becomes more obvious as the dataset becomes bigger.

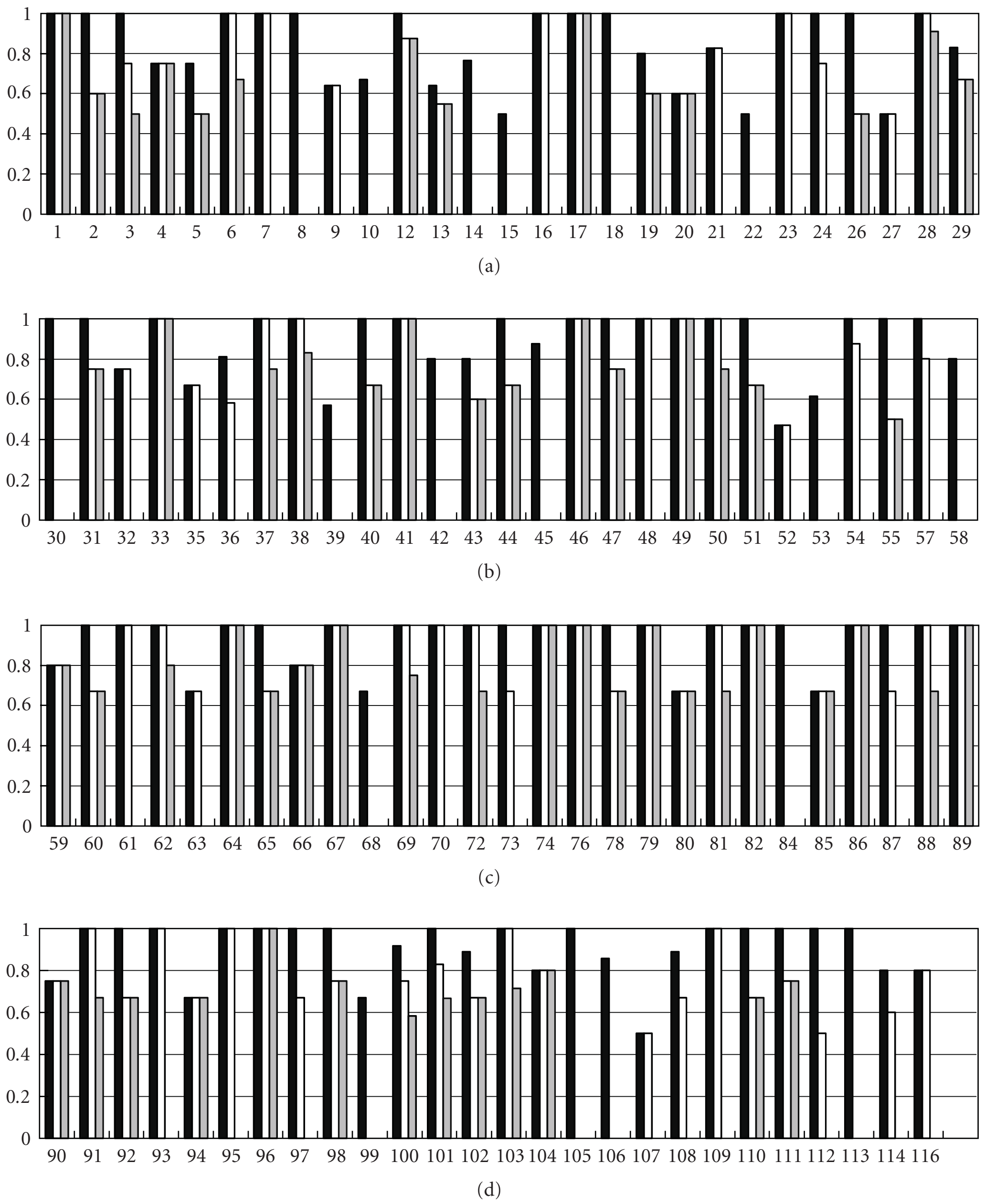


Figure 4: The functional coherence in each of the 116 groups, computed as the ratio of the number of proteins having a specific functional category to the number of proteins in the group. The black, white, and grey bars represent functional categories with the ratios ≥ 0.5 and the maximum number of such ratios is limited to 3 in each group.

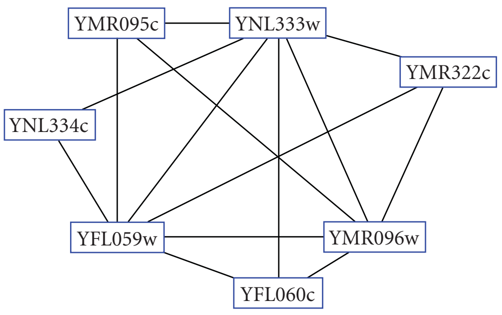


Figure 5: Group 93 identified as a near-clique by our algorithm.

## 4. CONCLUSION

Identifying hidden topological structures of protein interaction networks often unveil biologically relevant functional groups and structural complexes. We developed an efficient heuristic algorithm for finding cliques and near-cliques in protein interaction networks. From the interaction data of yeast proteins, the algorithm identified 116 near-cliques. Comparison with the experimental data showed that 59% of the near-cliques have at least one function shared by all the proteins within a near-clique, and that 56% of the near-cliques show a good agreement with known protein

Table 3: The near-cliques matched with experimentally determined protein complexes cataloged in MIPS. The overlap column represents the number of proteins common to the near-cliques and the protein complexes.

| Group ID | Group size | MIPS tag of protein complex | Protein complex size | Overlap (common proteins) | $\ln(P_{\text{overlap}})$ | Main functional categories |
|---|---|---|---|---|---|---|
| 3 | 4 | 295 | 2 | 2 | −14.28 | 42.10/43.01.03.09 |
| 5 | 4 | 510.190.110 | 13 | 3 | −16.32 | 01.03.16.01 |
| 6 | 6 | 320 | 8 | 4 | −23.40 | 11.02.03.04/10.01.09.05/16.03.01 |
| 9 | 22 | 550.1.149 | 88 | 21 | −84.60 | 11.04.01/12.01 |
| 12 | 8 | 550.1.148 | 35 | 8 | −39.49 | 11.04.03.05/16.03.03/11.04.03.01 |
| 13 | 11 | 550.1.7 | 10 | 8 | −47.55 | 10.03.01/14.01/16.01 |
| 14 | 17 | 140.30 | 32 | 11 | −46.65 | 10.03.01 |
| 16 | 4 | 550.1.44 | 9 | 3 | −17.55 | 14.07.02.02/01.05.01.03.02.02 |
| 17 | 3 | 410.40.20 | 3 | 3 | −23.36 | 16.03.01 /16.03.01 /10.01.05.01/10.01.03.05 /10.01.03.01 |
| 18 | 4 | 440.30.30 | 11 | 4 | −24.56 | 11.04.02 |
| 21 | 23 | 470.20 | 5 | 5 | −26.71 | 43.01.03.05/34.11.03.07 |
| 23 | 4 | 510.160 | 4 | 4 | −30.36 | 11.02.03.04.01/01.05.04 |
| 24 | 4 | 360.10 | 36 | 4 | −19.38 | 14.13.01.01/14.07.11 |
| 25 | 36 | 550.1.138 | 36 | 11 | −34.43 | none |
| 26 | 4 | 550.2.317 | 3 | 3 | −21.98 | 20.09.07.03/20.09.07.05/14.10 |
| 27 | 10 | 130 | 8 | 4 | −20.77 | 42.04/16.01 |
| 28 | 11 | 60 | 11 | 11 | −74.71 | 14.13.01.01/10.03.01.01.11/14.07.05 /14.10 /16.01 /16.19.03 |
| 29 | 6 | 120.20 | 4 | 4 | −27.65 | 43.01.03.05/40.01/34.07.01/34.01/32.01.09 /11.02.02/10.03.01.01.09/10.03.01.01.03 |
| 30 | 8 | 550.1.142 | 25 | 4 | −16.68 | 12.04 |
| 31 | 4 | 140.30.30.30 | 3 | 2 | −13.18 | 10.03.01/42.04/20.09 |
| 32 | 3 | 510.20 | 4 | 3 | −21.98 | 12.04.02/12.01.01 |
| 36 | 31 | 230.20.20 | 16 | 14 | −67.99 | 11.02.03.04/10.01.09.05 |
| 37 | 4 | 410.30 | 16 | 4 | −22.85 | 16.03.01/10.01.03.03/10.01.03.01/16.19.03 |
| 38 | 6 | 550.1.81 | 7 | 5 | −32.29 | 43.01.03.05/20.09.16.09.03/16.01/20.09.07.27 /10.03.03 |
| 40 | 3 | 410.30 | 16 | 3 | −17.03 | 10.01.03.03/10.01.09.05 /11.02.03.04/16.19.03 /34.11.03.07 |
| 44 | 3 | 350.10.10 | 2 | 2 | −14.97 | 14.10/01.04.01/14.13/16.19.03/20.01.10/20.09.04 |
| 45 | 8 | 500.10.40 | 7 | 6 | −38.44 | 12.04.01 |
| 47 | 4 | 410.40.30 | 5 | 3 | −19.67 | 10.01.05.01/10.01.03.05/10.03.01.03/16.03.01 /18.02.01 |
| 48 | 3 | 410.40.90 | 3 | 3 | −23.36 | 10.01.05.01/10.01.03.05 |
| 49 | 6 | 290.20.10 | 5 | 5 | −35.34 | 20.09.04/20.01.10/14.04/42.16 |
| 50 | 4 | 550.1.29 | 16 | 4 | −22.85 | 32.01/2.19/01.05.01.03.01/01.05.01.01.01 |
| 52 | 3 | 550.3.82 | 2 | 2 | −14.97 | 12.04.03/01.03.16.01 |
| 53 | 13 | 260.30.20 | 11 | 6 | −30.15 | 20.09.07.03 |
| 54 | 8 | 510.40.10 | 13 | 7 | −40.63 | 11.02.03.01/11.02.03.04 |
| 58 | 5 | 550.2.436 | 2 | 2 | −13.77 | 20.09.18 |
| 59 | 5 | 470.10 | 6 | 5 | −35.34 | 01.04.01 /01.05.04/14.07.03/30.01.05 /43.01.03.05/11.02.03.04 |

Table 3: Continued.

| Group ID | Group size | MIPS tag of protein complex | Protein complex size | Overlap (common proteins) | $\ln(P_{\text{overlap}})$ | Main functional categories |
|---|---|---|---|---|---|---|
| 61 | 5 | 440.12.10 | 7 | 5 | −34.08 | 11.04.01/01.03.16.01 |
| 64 | 3 | 400 | 10 | 3 | −18.57 | 43.01.03.09/01.06.01.07.11 |
| 66 | 5 | 550.1.212 | 35 | 5 | −24.44 | 16.19.03/10.01.05.01/01.04.01/16.03.01 |
| 67 | 4 | 510.190.40 | 5 | 4 | −28.75 | 11.04/01.03.16.01/11.02.03.04/34.11.03.07 |
| 70 | 3 | 440.12.30 | 3 | 3 | −23.36 | 11.06.01/11.04.01 |
| 72 | 3 | 550.1.84 | 7 | 3 | −19.80 | 20.09.13/16.01/14.10/20.09.18 |
| 74 | 3 | 510.180.30.10 | 2 | 2 | −14.97 | 42.10.03/32.01.09/16.03.03/10.03.01.03 /10.01.09.05/10.01.05.01 |
| 76 | 7 | 290.10 | 9 | 7 | −46.58 | 20.09.04/20.01.10/14.04/20.03 |
| 80 | 3 | 550.2.527 | 2 | 2 | −14.97 | 30.01.05.01/18.02.01/18.01.01/14.07.03 |
| 81 | 3 | 260.90 | 6 | 3 | −20.36 | 42.04.03/20.09.14/16.07/14.10/16.01/43.01.03.05 |
| 82 | 4 | 260.80 | 4 | 4 | −30.35 | 42.25/20.09.13/20.09.07/20.09.07/20.09.16.09.03 |
| 86 | 4 | 510.180.20 | 7 | 4 | −26.80 | 34.11.03.07/10.01.05.03.01/10.01.05.01/10.03.02 |
| 87 | 3 | 550.1.74 | 4 | 3 | −21.97 | 20.09.07/42.04.03 |
| 89 | 3 | 550.2.321 | 6 | 3 | −20.36 | 16.19.03 /14.13.01.01/14.07.11/01.04.01 |
| 91 | 3 | 550.2.317 | 3 | 3 | −23.36 | 16.03.01/11.02.01/11.02.02 |
| 95 | 3 | 90.30 | 2 | 2 | −14.97 | 34.01.01.03/14.10 |
| 96 | 4 | 110 | 4 | 4 | −30.35 | 43.01.03.05/11.02.03.04/16.19.01/30.01.09.07 /14.07.03/01.04.01 |
| 97 | 3 | 260.30.30.20 | 2 | 2 | −14.97 | 14.04 /20.09.07.05 |
| 100 | 12 | 140.20.30 | 7 | 4 | −20.60 | 43.01.03.05/42.04/40.01 |
| 101 | 6 | 550.3.12 | 7 | 4 | −24.09 | 10.03.01.01.03/43.01.03.05/30.01.05.01/14.07.03 |
| 102 | 9 | 445.20 | 4 | 4 | −25.52 | 16.01/16.19.03/10.03.01.01.03/14.13.01.01/14.10 /14.07.05 |
| 103 | 7 | 140.20.20 | 25 | 5 | −23.21 | 43.01.03.05/40.01/16.01/42.04.03/20.09.18.09.01 |
| 104 | 5 | 140.20.30 | 7 | 4 | −25.19 | 43.01.03.05/42.01/32.01.03/20.09.18.09.01 |
| 105 | 11 | 260.60 | 10 | 10 | −66.33 | 20.09.07.03 |
| 106 | 7 | 270.20.40 | 4 | 3 | −18.42 | 10.03.04.03 |
| 108 | 9 | 133.10 | 10 | 7 | −41.79 | 10.03.01.01/18.02.01 |
| 110 | 3 | 140.30.30 | 14 | 3 | −17.46 | 20.09.14/10.03.05.01/10.03.04.09/10.03.01.01.11 /02.45.11 |
| 112 | 5 | 270.20.20 | 3 | 3 | −21.05 | 10.03.04.05/16.01 |
| 113 | 5 | 270.20.10 | 3 | 3 | −21.05 | 10.03.04.05 |

Table 4: Comparison of our method and CFinder in terms of the number of functional categories shared by all the proteins in the groups.

| Functional categories common to all proteins in the group | Groups found by our program | Groups found by CFinder |
|---|---|---|
| < 1 | 9 (7.8%) | 35 (18.0%) |
| = 1 | 19 (16.3%) | 37 (18.9%) |
| = 2 | 27 (23.3%) | 35 (18.0%) |
| > 2 | 61 (52.6%) | 88 (45.1%) |
| Total | 116 (100%) | 195 (100%) |

Table 5: Running times of the programs on 3 data sets of yeast protein interactions on a Pentium IV 3.0 GHz processor with 512 MB memory.

| Program | MIPS data (4874 nodes 15660 edges) | DIP data (4932 nodes 17491 edges) | BOND data (18692 nodes 59516 edges) |
|---|---|---|---|
| Our program | 10.06 s | 13.06 s | 1 m 06.63 s |
| CFinder | 17.45 s | 24.33 s | 2 m 22.46 s |

MIPS data: PPI_180105.tab from MIPS (1/18/2005).

DIP data: yeast20071104.lst from DIP (11/04/2007).

BOND data: data from BOND (11/9/2007).

complexes,which are cataloged in the MIPS Saccharomyces cerevisiae genome database.

## ACKNOWLEDGMENTS

This work was supported by the Korea Science and Engineering Foundation (KOSEF) under Grant no. F01-2007-000-10140-0 and Grant no. F01-2007-000-10140-0 through the Systems Bio-Dynamics Research Center.

## REFERENCES

[1] D. Bu, Y. Zhao, L. Cai, et al., "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[2] R. Battiti and M. Protasi, "Reactive local search for the maximum clique problem," *Algorithmica*, vol. 29, no. 4, pp. 610–637, 2001.

[3] K. Katayama, A. Hamamoto, and H. Narihisa, "An effective local search for the maximum clique problem," *Information Processing Letters*, vol. 95, no. 5, pp. 503–511, 2005.

[4] B. Adamcsek, G. Palla, I. Farkas, I. Derényi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, 2006.

[5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[6] B.-H. Ju and K. Han, "Complexity management in visualizing protein interaction networks," *Bioinformatics*, vol. 19, supplement 1, pp. i177–i179, 2003.

[7] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[8] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12126, 2003.

[9] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[10] P. Uetz, L. Giot, G. Cagney, et al., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[11] A. Ruepp, A. Zollner, D. Maier, et al., "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes," *Nucleic Acids Research*, vol. 32, no. 18, pp. 5539–5545, 2004.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 371710, 9 pages
doi:10.1155/2008/371710

*Research Article*

# Detection of Gene Interactions Based on Syntactic Relations

**Mi-Young Kim**

*School of Computer Science and Engineering, Sungshin Women's University, Seoul 136-742, Korea*

Correspondence should be addressed to Mi-Young Kim, miykim@sungshin.ac.kr

Received 29 August 2007; Accepted 19 December 2007

Recommended by Daniel Howard

Interactions between proteins and genes are considered essential in the description of biomolecular phenomena, and networks of interactions are applied in a system's biology approach. Recently, many studies have sought to extract information from biomolecular text using natural language processing technology. Previous studies have asserted that linguistic information is useful for improving the detection of gene interactions. In particular, syntactic relations among linguistic information are good for detecting gene interactions. However, previous systems give a reasonably good precision but poor recall. To improve recall without sacrificing precision, this paper proposes a three-phase method for detecting gene interactions based on syntactic relations. In the first phase, we retrieve syntactic encapsulation categories for each candidate agent and target. In the second phase, we construct a verb list that indicates the nature of the interaction between pairs of genes. In the last phase, we determine direction rules to detect which of two genes is the agent or target. Even without biomolecular knowledge, our method performs reasonably well using a small training dataset. While the first phase contributes to improve recall, the second and third phases contribute to improve precision. In the experimental results using ICML 05 Workshop on Learning Language in Logic (LLL05) data, our proposed method gave an F-measure of 67.2% for the test data, significantly outperforming previous methods. We also describe the contribution of each phase to the performance.

## 1. INTRODUCTION

Determining interactions between proteins and genes are essential in describing biomolecular phenomena [1]. Thus, many recent studies have sought to extract interaction information from biomolecular text using natural language processing technology. However, we have insufficient biomolecular data annotated with linguistic information. In 2005, the ICML05 Workshop on Learning Language in Logic (LLL05) task provided a small training dataset annotated with POS-tags and syntactic relations. This was an experimental challenge for gene interactions using linguistic information. Previous studies have insisted that linguistic information was useful for improving the detection of gene interactions. However, the experimental results for the LLL05 data gave a reasonable precision but poor recall. To improve recall without sacrificing precision, we propose a three-phase method to detect gene interactions using syntactic relation information, and apply it to a small training dataset lacking domain knowledge. Through experimentation, we show that our proposed method significantly outperforms existing methods, and describe the contribution of each phase to its performance.

This paper is organized as follows. Section 2 presents previous work on gene interactions. Section 3 explains our three-phase method in detail. Section 4 describes the training and test data used for our experiments and presents experimental results that demonstrate that our three-phase method is effective for detecting gene interactions. Finally, we provide our conclusions.

## 2. PREVIOUS WORK

The task of relation mining in the biomedical domain has been studied extensively in recent years. Current research includes protein-protein interactions [2, 3], subcellular locations [4], and disease-treatment relationships [5], and systems based on sequence modeling and pattern- or rule-based extraction best detect protein-protein interactions [2, 6, 7]. Using text mining technology for automatic protein(gene) interactions resulted in high precision, but low recall [8]. Many studies have used linguistic information to improve performance in detecting gene interactions. To improve recall without sacrificing precision, Otasek et al. [8] expanded the diversity of sentence structures recognized by a syntactic parser through additional training, and Park et al. [9] presented a method using bidirectional incremental parsing. Experiments deduced 182 relations out of 492 sentences showing 48% recall and 80% precision. Many linguistic processes have been used to deduce gene interactions, including bidirectional incremental parsing, combinatory categorical grammar (CCG), coordination, apposition, compound noun processing, and positive/negative predicate learning. With these methods, linguistic information achieved reasonable precision, but still poor recall.

Blaschke et al. [10] assumed that sentences derived from sets of abstracts contained a significant number of protein names connected by verbs that indicate the type of relationship between them. They restricted the problem domain and imposed several strong assumptions that included prespecified protein names and a limited set of verbs to represent actions. Consequently, they constructed simple verb rules only for six proteins.

Several works examining gene interactions are based on LLL05 open data. Hakenberg et al. [11] used sentence alignment and finite-state automata optimized with a genetic algorithm. First, they applied a pattern-generating algorithm. Then, they learned patterns with finite-state automata based on a genetic algorithm. For example, "Agent1, Target3, Pattern2" implies that Agent1 interacts with Target3 via Pattern2. In biomolecular text, the agent or target can be encapsulated in another term based on some conditions, for example, apposition, modifying nouns, and so on. However, the method in [11] cannot deal with a situation in which genes are encapsulated in other terms via syntactic relations. They did not use linguistic information provided in the LLL05 data. Error analysis revealed that they wrongly detected an agent and its target in a pair of genes, although they correctly detected two genes that interact with each other. Linguistic information might correct this type of error.

Greenwood et al. [12] extracted patterns based on paths in MINIPAR dependency trees [13]. The nodes in the dependency trees from which patterns were derived were either a lexical item or a semantic category, such as a gene, protein, agent, or target. Patterns were learned using a weakly supervised bootstrapping method. They extended the patterns based on eight seed patterns and trained the model using the basic dataset without coreference, as provided by the LLL05 challenge organizers. The F-measure for the test data in LLL05 was 14.8%. The failure of the system to extract meaningful relations can be traced back to the errors that MINIPAR introduced in the dependency trees.

Goadrich et al. [14] used Gleaner as an inductive logic programming approach and further applied Brill Tagger, a shallow parser based on conditional random fields, and Porter stemmer. They also used much linguistic information, including sentence-structure predicates, the frequencies of words, lexical properties, and semantic knowledge using Mesh. The F-measure for the test data was 25.1%. Gleaner suffered from not distinguishing between an agent and a target well because no syntactic structure was used.

Popelinsky and Blatak [15] used Brill Tagger and WordNet, and Katrenko et al. [16] created a simple ontology specifically for use in the LLL05 challenge. However, they did not show reasonable recall.

Riedel and Klein [17] obtained the best performance on the LLL05 challenge task using syntactic chains. They assumed that clauses had to connect both genes transitively. Therefore, they generated a set of clauses based on chains of syntactic relations between two genes. The method achieved an F-measure of 52.6% on the dataset without coreferences, demonstrating that using syntactic information from the annotated datasets significantly improved performance. A CCG parser handled both POS-tagging and parsing. However, recall was only 46.2%, and the system needs to improve recall.

For GENIA and ATCR data, Rinaldi et al. [18] also used linguistic approach. They find agents and targets from the syntactic patterns directly connected with interaction verbs with subject or object functions. So, they do not consider the case that agent or target is encapsulated in another term, and indirectly connected with interaction verbs. In addition, there is a limit that they find agents and targets only from the subject and object relations.

Combining syntactic dependency information with features based on word sequences could lead to further improvements in performance, as demonstrated by the more recent approaches to relation extraction [19–21].

We build on the conclusion of the previous work that linguistic information, especially syntactic information, is an important key for detecting gene interactions. However, we need a more robust method to improve recall without sacrificing precision. Based on syntactic relation information, we propose a three-phase-based method for detecting gene interactions.

Greenwood et al. [12] mentioned the failure of the system to extract meaningful relations can be traced back to the errors of the applied syntactic analyzer. If we use the annotated LLL05 syntactic relation information, we cannot testify the robustness of our system in real time. So, we also experiment the performance of our system based on a real-syntactic analyzer.

To objectively compare the performance of our system with that of previous systems, we use LLL05 data. In the next section, we explain our proposed three-phase method in detail.

## 3. THREE-PHASE DETECTION OF GENE INTERACTIONS

Let us explain LLL05 data formats. The LLL05 challenge focuses on extracting information on gene interactions in *Bacillus subtilis*. The training dataset is decomposed into two subsets of increasing difficulty. The first subset does not include coreferences or ellipsis, unlike the second subset. The training set without coreferences consists of 55 sentences, including 106 examples of genic interactions. It contains 70 examples of action, 30 examples of binding and promoter, and 6 examples of regulation.

A syntactic relation is important linguistic information for detecting the structure of text. Algorithm 1 shows one

| | |
|---|---|
| **ID** | 11064201-3 |
| **sentence** | In this mutant, expression of the spoIIG gene, whose transcription depends on both sigma(A) and the phosphorylated Spo0A protein, Spo0A∼P, a major transcription factor during early stages of sporulation, was greatly reduced at 43 degrees C. |
| **words** | word(0,“In,”0,1) word(1,“this,”3,6) word(2,“mutant,”8,13) word(3,“expression,”16,25) word(4,“of,”27,28) word(5,“the,”30,32) word(6,“spoIIG,”34,39) word(7,“gene,”41,44) word(8,“whose,”47,51) word(9,“transcription,”53,65) word(10,“depends,”67,73) word(11,“on,”75,76) word(12,“both,”78,81) word(13,“sigma(A),”83,90) word(14,“and,”92,94) word(15,“the,”96,98) word(16,“phosphorylated,”100,113) word(17,“Spo0A,”115,119) word(18,“protein,”121,127) word(19,“Spo0A∼P,”130,136) word(20,“a,”139,139) word(21,“major,”141,145) word(22,“transcription,”147,159) word(23,“factor,”161,166) word(24,“during,”168,173) word(25,“early,”175,179) word(26,“stages,”181,186) word(27,“of,”188,189) word(28,“sporulation,”191,201) word(29,“was,”204,206) word(30,“greatly,”208,214) word(31,“reduced,”216,222) word(32,“at,”224,225) word(33,“43,”227,228) word(34,“degrees,”230,236) word(35,“C,”238,238) |
| **lemmas** | lemma(0,“in”) lemma(1,“this”) lemma(2,“mutant”) lemma(3,“expression”) lemma(4,“of”) lemma(5,“the”) lemma(6,“spoIIG”) lemma(7,“gene”) lemma(8,“whose”) lemma(9,“transcription”) lemma(10,“depend”) lemma(11,“on”) lemma(12,“both”) lemma(13,“sigA”) lemma(14,“and”) lemma(15,“the”) lemma(16,“phosphorylated”) lemma(17,“spo0A”) lemma(18,“protein”) lemma(19,“Spo0A-P”) lemma(20,“a”) lemma(21,“major”) lemma(22,“transcription”) lemma(23,“factor”) lemma(24,“during”) lemma(25,“early”) lemma(26,“stage”) lemma(27,“of”) lemma(28,“sporulation”) lemma(29,“be”) lemma(30,“greatly”) lemma(31,“reduce”) lemma(32,“at”) lemma(33,“43”) lemma(34,“degree”) lemma(35,“C”) |
| **syntactic_relations** | relation(“subj:V_PASS-N,”31,3) relation(“mod_att:N-N,”7,6) relation(“mod_att:N-ADJ,”34,33) relation(“comp_during:N-N,”23,26) relation(“comp_of:N-N,”26,28) relation(“comp_on:V-N,”10,13) relation(“mod_att:N-N,”23,22) relation(“mod_att:N-ADJ,”18,16) relation(“mod:V_PASS-ADV,”31,30) relation(“mod_att:N-ADJ,”26,25) relation(“mod_att:N-ADJ,”23,21) relation(“mod_att:N-N,”18,17) relation(“comp_on:V-N,”10,18) relation(“appos,”19,23) relation(“subj:V-N,”10,9) relation(“appos,”18,19) relation(“comp_of:N-N,”3,7) relation(“comp_in:V-N,”31,2) relation(“comp_of:N-N,”9,7) relation(“comp_at:V_PASS-N,” 31,34) relation(“mod_att:N-N,”34,35) |
| **agents** | agent(13) agent(17) |
| targets | target(6) |
| **genic_relations** | genic_interaction(13,6) genic_interaction(17,6) |

Algorithm 1: Example of LLL05 training data.

example of syntactic relations between two genes in the LLL05 data. The syntactic relations provided in LLL05 were of the form *relation*(*rel*$_i$, $w_i$, $w_j$), where *rel*$_i$ is one of a fixed set of syntactic relations between $w_i$ and $w_j$ assigned by the LLL parser. The detailed contents about LLL05 training data are described in Algorithm 2.

Figure 1 shows an example of a syntactic path. In Figure 1, **Spo0A**(agent) goes through four terms to reach **spoIIG**(target). The chain of terms is ⟨**Spo0A**(agent) → protein(N) → depend(V) → transcription(N) → gene(N) → **spoIIG**(target)⟩. In the chain, node depend(V) is the verb that indicates the interaction between **Spo0A**(agent) and **spoIIG**(target). However, depend(V) has direct syntactic relations with protein(N) and transcription(N), not with **Spo0A**(agent) or **spoIIG**(target). In other words, **Spo0A**(agent) was encapsulated in protein(N) with the relation (“mod_att”), and **spoIIG**(target) was encapsulated in transcription(N) with the relation (“mod_att”) and (“comp_of”).

Without any domain knowledge of biomolecular text, we automatically detect gene interactions using syntactic relations annotated in the LLL05 data. In the first phase, to improve recall, we detect the relations that encapsulate an agent or target. In the second phase, we automatically extract “interaction verbs” that indicate interactions between two genes. Next, to improve precision, we must determine which of the two genes is the agent and which is the target. To determine the agent and target for two genes, we learn direction rules on the relations from agent to target in the third phase. The three phases are explained in detail from the next subsection.

```
1> ID                      : unique identifier

2> sentence                : the original sentence
3> words                   : list of the sentence words
                           transcription-word (id_word, "string_word," start_word postion, end_word position)
                           ex> word(0,"Both,"0,3)

4> lemmas                  : normalized form of a word
                           transcription-lemma(id_word, "string_lemma")
                           ex> lemma(0,"Both")

5> syntactic_relations     : syntactic relation between two words
                           transcription-relation("string_relation," id of the head, id of the dependent)
                           (a) string_relation is expressed as
                               "syntactic category:POStag of the head-POStag of the dependent."
                               ex> relation("mod_att:N-N,"8,7)
                           (b) POS-tags: V, V_PASS, N, ADJ, ADV
                           (c) syntactic categories: APPOS, COMP_prep, MOD, MOD_ATT, MOD_POST,
                               MOD_PRED, NEG, OBJ, SUBJ
6> agents                  : agent of the genic interaction
                           transcription-agent(id of the word)

7> targets                 : target of the genic interaction
                           transcription-target (id of the word)

8> genic_interactions      : an interaction between an agent and a target
                           transcription-genic_interaction(id of the agent, id of the target)

Please see http://genome.jouy.inra.fr/texte/LLLchallenge
```

Algorithm 2: Detailed contents of LLL05 training data.

### 3.1. Phase 1: constructing syntactic encapsulation categories for agents and targets

An agent or target gene is usually encapsulated in another term, and the verb that indicates the interaction between two genes has syntactic relations with two terms that encapsulate the genes. To improve recall for gene interactions, we must detect the encapsulation categories for candidate agents and targets. First, we find the syntactic chain from an agent to its target. In Figure 1, depend(V) is the verb that indicates an interaction between **Spo0A**(agent) and **spoIIG**(target). In this paper, we call the verb that indicates the interaction between an agent and its target an "interaction verb." As mentioned above, depend(V) has syntactic relations with protein(N) and transcription(N), but not with **Spo0A**(agent) or **spoIIG**(target). In a syntactic chain from an agent to its target, we call the node preceding an interaction verb a "metaagent," and the node following an interaction verb a "metatarget." In Figure 1, protein(N) is a metaagent, and transcription(N) is a metatarget.

We define the syntactic categories connecting an agent(target) and a metaagent(metatarget) "syntactic encapsulation categories." In Figure 1, mod_att and comp_of are examples of the syntactic encapsulation categories. To detect a metaagent and a metatarget, we should first identify an interaction verb in a syntactic chain. However, in the automatically obtained syntactic chains, we do not know which verb is an interaction verb. To overcome the problem, we extract the syntactic encapsulation categories from the syntactic chains that include only one verb in the training dataset.

### 3.2. Phase 2: extracting interaction verbs that indicate an interaction between two genes

To detect gene interactions, we must recognize the interaction verbs. In the second phase, we retrieve the interaction verbs that indicate an interaction between two genes. The verbs can be extracted while the first phase is performed. If we consider only the syntactic chains that contain only one verb, the size of the interaction verbs becomes very small. Since the LLL05 training dataset is small, we collect all the verbs in the syntactic chains from an agent to its target.

### 3.3. Phase 3: learning direction rules for detecting the agent and target in a pair of genes

According to the first and second phases, we can detect two genes that interact with each other.

Previous studies made many errors in attempts to recognize which of two genes was the agent or target. The incorrect detection of an agent and a target results in low precision. Therefore, a new method is required to recognize an agent and its target correctly in a pair of genes. In the third phase, we propose learning the directions of the syntactic relations in the syntactic path from an agent to its target. If we do not

Example sentence:
⟨In this mutant, expression of the spoIIG gene, whose transcription depends on both sigma(A) and the phosphorylated Spo0A protein, Spo0A~P, a major transcription factor during early stages of sporulation, was greatly reduced at 43 degrees C.⟩

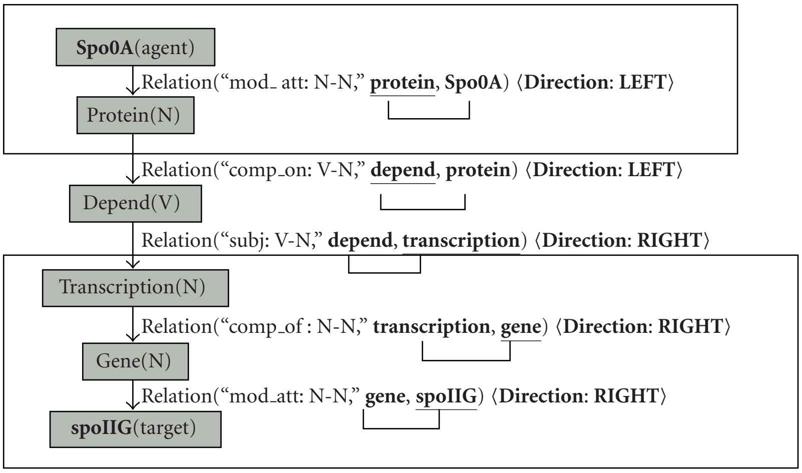

**Spo0A**(agent)
Relation("mod_ att: N-N," **protein**, **Spo0A**) ⟨**Direction**: **LEFT**⟩
Protein(N)
Relation("comp_on: V-N," **depend**, **protein**) ⟨**Direction**: **LEFT**⟩
Depend(V)
Relation("subj: V-N," **depend**, **transcription**) ⟨**Direction**: **RIGHT**⟩
Transcription(N)
Relation("comp_of : N-N," **transcription**, **gene**) ⟨**Direction**: **RIGHT**⟩
Gene(N)
Relation("mod_att: N-N," **gene**, **spoIIG**) ⟨**Direction**: **RIGHT**⟩
**spoIIG**(target)

Figure 1: Example of a syntactic path based on LLL05 annotations.

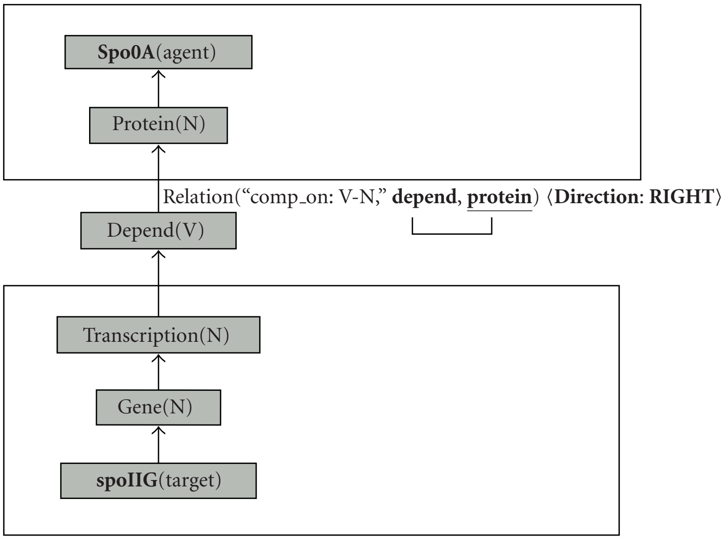

**Spo0A**(agent)
Protein(N)
Relation("comp_on: V-N," **depend**, **protein**) ⟨**Direction**: **RIGHT**⟩
Depend(V)
Transcription(N)
Gene(N)
**spoIIG**(target)

Figure 2: Reverse syntactic path of Figure 1 for a negative rule.

permit the reverse direction, the agent and target will not be detected wrongly and thus improve the precision.

We learn the direction of a syntactic relation related with an interaction verb. For a syntactic relation, direction is defined as follows. If a syntactic relation is *relation(syntactic category, current node, next node),* the direction is "RIGHT," since the next node is written to the right of the current node. If a syntactic relation is *relation(syntactic category, next node, current node),* the direction is "LEFT" because the next node is written to the left of the current node. Figure 1 also shows an example of direction information of a syntactic path. Among the directions, we retrieve only the direction information of an interaction verb.

The direction information is dependent on the syntactic category of the relation and the lexical word of the current node. In learning, we retrieve a syntactic category (a lexical word) and direction information for an interaction verb, and we make a template ⟨lexical word, syntactic category, direction⟩.

We construct direction information for all relations concerning interaction verbs in the training data. Based on the direction information, we learn direction rules. Let us explain the direction rule-learning algorithm, which is shown in Algorithm 3.

We obtain two types of rule set. One is a positive rule set obtained by learning the direction from an agent to its target.

```
1> Alignment of a positive rule set
    (1.1) Collect ⟨lexical word A, relation B, direction C⟩ in the paths from all Agents to their Targets.
    (1.2) For any lexical word A, and relation B, if both of ⟨A, B, RIGHT⟩ and ⟨A, B, LEFT⟩ exist in
          the positive rule set, we remove both rules, and add a modified rule ⟨A, B, ANY⟩.
2> Alignment of a negative rule set
    (2.1) Collect ⟨lexical word A, relation B, direction C⟩ in the paths from all Targets to their Agents.
    (2.2) For any lexical word A, and relation B, if both of ⟨A, B, RIGHT⟩ and ⟨A, B, LEFT⟩ exist in
          the negative rule set, we remove both rules, and add a modified rule ⟨A, B, ANY⟩.

3> Construction of Final direction rules
    For every rule ⟨A, B, C⟩ in the positive rule set, for any lexical word A, relation B, and direction C.

        (3.1) If ⟨A, B, C⟩ also exists in the negative rule set, we obtain a direction rule ⟨A, B, ANY⟩.
        (3.2) Otherwise, if ⟨A, B, OPPOSITE C⟩ exists in the negative rule set, we obtain a direction rule
              ⟨A, B, C⟩.
        (3.3) Otherwise, we obtain a direction rule ⟨A, B, C⟩.
```

Algorithm 3: Direction rule-learning algorithm.

Table 1: Positive and negative rule sets for the sentence in Figure 1.

| | |
|---|---|
| Positive rule | ⟨depend, subj, RIGHT⟩ |
| Negative rule | ⟨depend, comp_on, RIGHT⟩ |

The other is a negative rule set obtained by learning the direction from a target to its agent in reverse order. Figure 2 shows the reverse syntactic path from a target to its agent of the sentence in Figure 1. The positive and negative rules for the sentence in Figure 1 are shown in Table 1. From the positive and negative rule sets, we construct direction rules according to the following subsections.

### *3.3.1. Alignment of positive/negative rule sets*

First, we align the positive and negative rule sets. Here, "align" means the modification of any conflict in a rule set. For any lexical word A and relation B, if a conflict of two direction rules exists in a rule set, then we remove both rules, and add a modified rule ⟨A, B, ANY⟩. Because the direction information is not trustworthy, we set direction "ANY." "ANY" means any direction is okay. The process for aligning a rule set is shown in 1> and 2> of Algorithm 3.

### *3.3.2. Construction of direction rules from positive and negative rule sets*

After alignment of positive and negative rule sets, we construct direction rules from the two rule sets. The algorithm used to obtain direction rules is shown in 3> of Algorithm 3.

Consider every rule ⟨A, B, C⟩ in the positive rule set, for any lexical word A and relation B, and direction C.

In Algorithm 3, (3.1) case indicates that direction information C is changed to ANY. Since the same direction exists in both the positive and negative rule sets, the direction information is not trustworthy. Therefore, we change the direction information into ANY.

In (3.2) case, the direction information C in the positive rule is still used in the obtained direction rule. The case indicates that the negative rule set has "OPPOSITE C" direction. If C is "RIGHT," then "OPPOSITE C" means "LEFT." Otherwise, if C is "LEFT," then "OPPOSITE C" means "RIGHT." Since the direction in the negative rule set is opposite with that in the positive rule set, the direction information in the template is trustworthy.

(3.3) case indicates that the negative rule set does not have any rule concerning A and B. The obtained direction rule is same with the original template in the positive rule set. The examples of learned direction rules are shown in Table 2. For an interaction verb A, the relations not learned in the training data can appear in the test data. So, we add a default rule ⟨A, otherwise, ANY⟩ as described in Table 2. The default rule permits any direction is okay for other relations not appearing in the training data. Because the training data is so small, the default rule can resolve data sparseness problem.

## *3.4. Applying our proposed method to test data*

The procedure to detect gene interactions in the test data is as follows. We detect agent candidates from the test set using the gene dictionary provided by LLL05. Starting from an agent candidate node, we extend all possible syntactic paths. The obtained syntactic encapsulation categories, interaction verbs, and direction rules through three phases are applied to test data according to the following procedure.

For each syntactic chain, we repeat the following procedure.

(1) If a current node is a gene and syntactic chain contains any interaction verb, then we determine that the current node is a target, and stop the extension of the syntactic chain.

(2) Otherwise, if the category of the syntactic relation of the next node candidate is a syntactic encapsulation

Table 2: Examples of direction rules learned through the third phase (based on MINIPAR).

| Lexical word | Relation | Direction | Lexical word | Relation | Direction |
|---|---|---|---|---|---|
| activate | aux | RIGHT | affect | aux | LEFT |
| activate | otherwise | ANY | affect | *i* | RIGHT |
| bind | conj | LEFT | affect | obj | ANY |
| bind | *i* | RIGHT | affect | otherwise | ANY |
| ... | ... | ... | drive | obj | LEFT |
| bind | otherwise | ANY | drive | *s* | RIGHT |

Table 3: Performances of our system and other previous systems using LLL05 syntactic tags.

| Performance on test data(%) | | Hakenberg et al. [11] | Goadrich et al. [14] | Riedel and Klein [17] | Popelinsky and Blatak [15] | Katrenko et al. [16] | Our system (Using LLL05 tags) |
|---|---|---|---|---|---|---|---|
| Using LLL05 syntactic tags | Precision | 28.1 | 28.3 | 60.9 | 46.5 | 39.2 | 67.9 |
| | Recall | 31.4 | 79.6 | 46.2 | 50.0 | 26.5 | 66.6 |
| | F-measure | 29.6 | 41.7 | 52.6 | 48.2 | 31.6 | 67.2 |

Table 4: Performances of our system and other previous systems using MINI-PAR.

| Performance on test data(%) | | Greenwood et al. [12] | Our system (Using MINIPAR) |
|---|---|---|---|
| Using MINIPAR | Precision | 22.2 | 32.4 |
| | Recall | 11.1 | 68.5 |
| | F-measure | 14.8 | 44.0 |

Table 5: Change in performance when one phase is removed.

| | Performance on the test data based on LLL05 syntactic relations(%) | |
|---|---|---|
| Using all phases | Precision | 67.9 |
| | Recall | 66.6 |
| | F-measure | 67.2 |
| Without the first phase (there is no "syntactic encapsulation categories") | Precision | 0 |
| | Recall | 0 |
| | F-measure | 0 |
| Without the second phase (all verbs are considered "interaction verbs") | Precision | 24.6 |
| | Recall | 88.8 |
| | F-measure | 38.5 |
| Without the third phase (there is no syntactic direction information) | Precision | 39.7 |
| | Recall | 72.2 |
| | F-measure | 51.3 |

category, we extend the syntactic chain by adding the next node candidate.

(3) Otherwise, if the current lexical word is an interaction verb and the direction of the next node candidate is consistent with the direction rules, then we extend the syntactic chain.

In the finally obtained syntactic chains, we determine that the first node is an agent and the last node is its target.

## 4. EXPERIMENTAL EVALUATION

### 4.1. *Performance of our three-phase method versus those of other methods*

With more and more biomedical datasets becoming publicly available, there has been some research effort on corpus design issues and usage in biomedical natural language processing [22, 23]. For a reasonable comparison with previous methods, we applied the training and test data from the LLL05 challenge task. As mentioned before, the LLL05 training dataset without coreference consists of 55 sentences, including 106 genic interactions, and the test data consist of 144 sentences.

Our experiment focused on the following three points.

(1) Based on the LLL05 syntactic tags, the performance of our three-phase method versus that of previous methods.
(2) Based on a real-syntactic analyzer, the performance of our three-phase method versus that of previous methods.
(3) The change in performance when each phase is removed.

In the experiments, we obtained the following five results.

(1) Our three-phase detection method for gene interactions achieved an F-measure of 67.2% using LLL05-annotated syntactic relations, and 44.0% using a real-syntactic analyzer (see Tables 3 and 4).
(2) Using LLL05 syntactic tags, our three-phase method achieved an improvement of 14.6% to 37.6% over previous methods (see Table 3).
(3) Our method significantly outperformed Greenwood et al. [12], which also used MINIPAR (see Table 4).

(4) When the second or third phase was removed, the precision became significantly worse (see Table 5).

(5) When the first phase was removed, there were no interaction results. It means the first phase is important for the improvement of recall (see Table 5).

As shown in Table 3, of the systems evaluated, our system performed the best with a precision of 67.9%, recall of 66.6%, and an F-measure of 67.2 percent.

### 4.2. Discussion of results

We will summarize the significance of each phase introduced in Section 3. As shown in Table 5, every phase is important for its performance. Without the first phase, if no syntactic relations are considered encapsulation categories, then no pairs of genes are generated. Only this result shows the decrease of recall among three results in Table 5. It demonstrates that the syntactic encapsulation categories contribute to the improvement of recall.

Without the second phase, if all the verbs are considered interaction verbs, the precision is very low, which results from the generation of too many wrong syntactic paths. Without the third phase, if we do not consider direction information, then the recall increases and the precision significantly decreases, which also result from the construction of many wrong syntactic paths.

The experiments prove that the second and third phases contribute to the improvement of precision, and the first phase to the improvement of recall. We conclude that all three phases are important for detecting gene interactions.

To experiment the robustness of our method in real time, we have used MINIPAR, an existing syntactic analyzer. The system based on annotated syntactic relations in LLL05 significantly outperforms that using MINIPAR. This is because of the errors in syntactic relations and POS-tags that MINIPAR produced.

## 5. CONCLUSION

To improve recall without sacrificing precision, this paper proposes a three-phase method for the automatic detection of gene interactions using syntactic relations. The proposed method does not require domain knowledge. To improve recall, in the first phase, we construct syntactic encapsulation categories of agent and target. In the second phase, we construct interaction verbs that connect pairs of genes that interact with each other. To improve precision, in the third phase, we learn direction information to detect which of the two genes is the agent or target. The experimental results show that our three-phase method performs significantly better than previous methods. Our method achieved a precision of 67.9%, a recall of 66.6%, an F-measure of 67.2% using LLL05 syntactic relations. We conclude that our proposed three-phase method is effective for detecting gene interactions. Furthermore, we demonstrated that every phase is important for performance.

In the future, we need to expand the size of the training dataset and experiment with a large dataset.

## ACKNOWLEDGMENT

This work was supported by the Sungshin Women's University research grant of 2007.

## REFERENCES

[1] P. Uetz and R. L. Finley Jr., "From protein networks to biological systems," *FEBS Letters*, vol. 579, no. 8, pp. 1821–1827, 2005.

[2] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604–3612, 2004.

[3] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo, "Extracting human protein interactions from MEDLINE using a full-sentence parser," *Bioinformatics*, vol. 20, no. 5, pp. 604–611, 2004.

[4] B. J. Stapley, L. A. Kelley, and M. J. Sternberg, "Predicting the sub-cellular location of proteins from text using support vector machines," in *Proceedings of the 7th Pacific Symposium on Biocomputing*, pp. 374–385, Lihue, Hawaii, USA, January 2002.

[5] B. Rosario and M. Hearst, "Classifying semantic relations in bioscience texts," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pp. 430–437, Barcelona, Spain, July 2004.

[6] J. Xiao, J. Su, G. Zhou, and C. Tan, "Protein-protein interaction extraction: a supervised learning approach," in *Proceedings of the 1st Symposium on Semantic Mining in Biomedicine (SMBM '05)*, pp. 51–59, Hinxton, Cambridgeshire, UK, April 2005.

[7] J. Saric, L. Jensen, R. Ouzounova, I. Rojas, and P. Bork, "Large-scale extraction of protein/gene relations for model organisms," in *Proceedings of the Symposiumon SemanticMining in Biomedicine*, p. 50, Hinxton, Cambridgeshire, UK, April 2005.

[8] D. Otasek, K. Brown, and I. Jurisica, "Confirming protein-protein interactions by text mining," in *Proceedings of the 6th SIAM Conference on Text Mining*, Bethesda, Md, USA, April 2006.

[9] J. C. Park, H. S. Kim, and J. J. Kim, "Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar," in *Proceedings of the 6th Pacific Symposium on Biocomputing (PSB '01)*, pp. 396–407, Mauna Lani, Hawaii, USA, January 2001.

[10] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, pp. 60–67, Heidelberg, Germany, August 1999.

[11] J. Hakenberg, C. Plake, U. Leser, H. Kirsch, and D. R. Schuhmann, "LLL05 challenge: genic interaction extraction-identification of language patterns based on alignment and finite state automata," in *Proceedings of the ICML05 Workshop on Learning Language in Logic (LLL '05)*, pp. 38–45, Bonn, Germany, August 2005.

[12] M. A. Greenwood, M. Stevenson, Y. Guo, H. Harkema, and A. Roberts, "Automatically acquiring a linguistically motivated genic interaction extraction system," in *Proceedings of the ICML05 Workshop on Learning Language in Logic (LLL '05)*, Bonn, Germany, August 2005.

[13] D. Lin, "Dependency-based evaluation of MINIPAR," in *Proceedings of the Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May 1998.

[14] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning to extract genic interactions using Gleaner," in *Proceedings of the ICML05 Workshop on Learning Language in Logic (LLL05)*, Bonn, Germany, August 2005.

[15] L. Popelinsky and J. Blatak, "Learning genic interactions without expert domain knowledge: comparison of different ILP algorithms," in *Proceedings of the ICML05 Workshop on Learning Language in Logic (LLL '05)*, Bonn, Germany, August 2005.

[16] S. Katrenko, M. S. Marshall, M. Roos, and P. Adriaans, "Learning biological interactions from Medline abstracts," in *Proceedings of ICML05 Workshop on Learning Language in Logic (LLL '05)*, Bonn, Germany, August 2005.

[17] S. Riedel and E. Klein, "Genic interaction extraction with semantic and syntactic chains," in *Proceedings of the ICML05 Workshop on Learning Language in Logic (LLL '05)*, Bonn, Germany, August 2005.

[18] F. Rinaldi, G. Schneider, K. Kaljurand, et al., "Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach," *Artificial Intelligence in Medicine*, vol. 39, no. 2, pp. 127–136, 2007.

[19] S. Zhao and R. Grishman, "Extracting relations with integrated information using kernel methods," in *Proceedings of the Association for Computational Linguistics*, pp. 419–426, Ann Arbor, Mich, USA, June 2005.

[20] M. Zhang, J. Zhang, J. Su, and G. Zhou, "A composite kernel to extract relations between entities with both flat and structured features," in *Proceedings of the Computational Linguistics and Association for Computational Linguistics (COLING-ACL '06)*, Sydney, Australia, July 2006.

[21] J. Jiang and C. Zhai, "A systematic exploration of the feature space for relation extraction," in *Proceedings of Human Language Technologies: The North American Chapter of the Association for Computational Linguistics (NAACLHLT '07)*, pp. 113–120, Rochester, NY, USA, April 2007.

[22] K. B. Cohen, L. Fox, P. V. Ogren, and L. Hunter, "Corpus design for biomedical natural language processing," in *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, pp. 38–45, Detroit, Mich, USA, June 2005.

[23] K. B. Cohen, L. Fox, P. V. Ogren, and L. Hunter, "Empirical data on corpus design and usage in biomedical natural language processing," in *Proceedings of the American Medical Informatics Association (AMIA '05)*, pp. 156–160, Washington, DC, USA, November 2005.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 841312, 6 pages
doi:10.1155/2008/841312

*Research Article*

# Structure-Based Inhibitors Exhibit Differential Activities against *Helicobacter pylori* and *Escherichia coli* Undecaprenyl Pyrophosphate Synthases

**Chih-Jung Kuo,[1, 2] Rey-Ting Guo,[1, 2] I-Lin Lu,[3] Hun-Ge Liu,[4] Su-Ying Wu,[3] Tzu-Ping Ko,[4] Andrew H.-J. Wang,[1, 2, 4] and Po-Huang Liang[1, 2, 4]**

[1] *Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan*
[2] *Institute of Biochemical Sciences, National Taiwan University, Taipei 106, Taiwan*
[3] *Division of Biotechnology and Pharmaceutical Research, National Health Research Institutes, Chu-Nan, Miaw-Li 350, Taiwan*
[4] *Institute of Biological Chemistry, Academia Sinica, Taipei 11529, Taiwan*

Correspondence should be addressed to Po-Huang Liang, phliang@gate.sinica.edu.tw

Received 21 September 2007; Accepted 27 December 2007

Recommended by Daniel Howard

*Helicobacter Pylori* colonizes the human gastric epithelium and causes diseases such as gastritis, peptic ulcers, and stomach cancer. Undecaprenyl pyrophosphate synthase (UPPS), which catalyzes consecutive condensation reactions of farnesyl pyrophosphate with eight isopentenyl pyrophosphate to form lipid carrier for bacterial peptidoglycan biosynthesis, represents a potential target for developing new antibiotics. In this study, we solved the crystal structure of *H. pylori* UPPS and performed virtual screening of inhibitors from a library of 58,635 compounds. Two hits were found to exhibit differential activities against *Helicobacter Pylori* and *Escherichia coli* UPPS, giving the possibility of developing antibiotics specially targeting pathogenic *H. pylori* without killing the intestinal *E. coli.*

## 1. INTRODUCTION

Undecaprenyl pyrophosphate synthase (UPPS) catalyzes consecutive condensation reactions of farnesyl pyrophosphate (FPP) with eight molecules of isopentenyl pyrophosphate (IPP) to form $C_{55}$ undecaprenyl pyrophosphate (UPP), which acts as a lipid carrier to mediate bacterial peptidoglycan biosynthesis [1, 2]. This enzyme belongs to a group of *cis*-prenyltransferases which catalyze *cis*-double bonds during IPP condensation reactions [3, 4]. UPPS was first cloned from *Micrococcus luteus* and *Escherichia coli*, and their amino acid sequences were found conserved among the *cis*-prenyltransferases, but totally different from those of the *trans*-prenyltransferases [5–7], implying different catalytic mechanism [8, 9].

*Helicobacter Pylori* is a pathogen which causes chronic inflammation in the stomach [10]. The infection may evolve to peptic ulcerations and gastric neoplasias. Due to its unusual ability to survive in stomach under the low pH condition via proton pumps, *H. Pylori* infection becomes wide spreading and accounts for the increased cases of stomach carcinogenesis [11]. Antibiotics, such as proton pump inhibitors (PPI), amoxicillin, and clarithromycin, are used to treat the infected patients. When failed, empirical quadruple therapy (PPI-bismuch-tetracyclin-metronidazole) is then used as the second-line therapy [12]. Since UPPS is essential for bacterial survival, it could possibly serve as a target for new antibiotics. Even though the complex structures of *E. coli* UPPS with the FPP substrate or with its analogue (farnesyl thiopyrophosphate, FsPP) and IPP have been obtained [9, 13], no UPPS structure-derived inhibitors have been reported so far. As shown in this study, we solved the crystal structures of *H. pylori* UPPS and performed structure-based inhibitor discovery. Two hits were discovered through computer virtual screening from 58,635 compounds, which exhibited different level of inhibition against *E. coli* and *H. pylori* UPPS.

## 2. MATERIALS AND METHODS

### 2.1. Overexpression of *H. pylori* UPPS

The gene encoding UPPS from the *H. pylori* (ATCC43504) genomic DNA was amplified by using polymerase chain reaction (PCR). The forward primer 5′-GGTATTGAGGGTCGCTTGGATAGCACTCTCAAA-3′ and reverse primer 5′-AGAGGAGAGTTAGAGCCCTAGCATTTTAATTCCCC-3′ were utilized in the PCR. The PCR product was purified from 0.8% agarose gel electrophoresis. The DNA product was ligated with pET-32Xa/LIC vector and transformed into *E. coli* BL21 (DE3) for protein expression as previously described for expressing *E. coli* UPPS [14].

The C234A mutant was prepared by using QuikChange Site-Directed Mutagenesis Kit in conjunction with the wild-type gene template in the pET32Xa/Lic vector. The mutagenic forward primer was 5′-CGCAAATTCGGGGAATTAAAA GCC TAGTGAGGCTCTAACTCT-3′. The procedure of mutagenesis utilized a supercoiled double-stranded DNA (dsDNA) vector with an insert of interest and two synthetic forward and backward primers containing the desired mutation. The mutation was confirmed by sequencing the entire UPPS mutant gene of the plasmid obtained from overnight culture. The correct construct was subsequently transformed to *E. coli* BL21(DE3) for protein expression. The procedure for protein purification followed our reported protocol [15]. Each purified mutant UPPS was verified by mass spectroscopic analysis and its purity (>95%) was checked by SDS-PAGE.

### 2.2. Crystallization and data collection

*H. pylori* C234A UPPS mutant was crystallized using the hanging drop method from Hampton Research (Laguna Niguel, Calif, USA) by mixing 2 μL of the UPPS solution (10 mg/mL in 25 mM Tris, 150 mM NaCl, pH 8.0) with 2 μL of the mother liquor (0.15 M KSCN, 15% PEG600, and 2% PEG5KMME), and equilibrating with 500 μL of the mother liquor. Within 4 days, crystals grew to dimensions of about 0.5 × 0.5 × 0.2 mm, and then the crystals were soaked with a cryoprotectant solution of 0.2 M KSCN, 30% PEG600, and 5% PEG5KMME for 1 day. The structure of the C234A *H. pylori* UPPS in complex with FsPP was obtained by soaking the crystals with cryoprotectant solution of 2.5 mM $MgCl_2$, 2.5 mM IPP, 2.5 mM FsPP, 0.15 M KSCN, 15% PEG600, and 2% PEG5KMME. However, only the pyrophosphate of FsPP was found in the complex structure. The X-ray diffraction datasets for the structures of the C234A UPPS mutant and the complex with FsPP were collected to 1.88 Å and 2.5 Å resolution, respectively. Data for the C234A UPPS crystals were collected at beam line BL17B2 of the National Synchrotron Radiation Research Center (NSRRC, Hsinchu, Taiwan). Data for the C234A UPPS complexed with FsPP were collected in house using a Rigaku MicroMax002 X-ray generator equipped with an *R*-Axis IV++ image plate detector. The diffraction data were processed using the programs of HKL and HKL2000 [16]. Statistics for the dataset are listed in Table 1. Prior to use in structural refinements, 5% randomly selected reflections were set aside for calculating $R_{\text{free}}$ as a monitor [17].

### 2.3. Structure determination and refinement

The crystal structure of C234A UPPS was determined by molecular replacement method using the *Crystallography & NMR System* (CNS) program [18]. The orthorhombic crystal contained one UPPS dimer in an asymmetric unit. The models of PDB 1V7U (*E. coli* UPPS structure bound with FPP, chain A) [13] were used as search model to yield a good resolution for the *H. pylori* UPPS. The space group was determined as $P2_12_12_1$. With all solvent and cofactor molecules removed, the model yielded an initial *R*-value of 0.50 using all positive reflections at 1.88 Å resolution upon rigid-body refinement.

The 2Fo-Fc difference Fourier map showed clear electron densities for most amino acid residues. The residues of catalytic loop of 58–67 in chain A, 56–71 and 150–158 in chain B were disordered. Subsequent refinement with incorporation of 581 water molecules according to 1.0 σ map level yielded *R* and $R_{\text{free}}$ values of 0.193 and 0.240, respectively, at 1.88 Å resolution. By employing similar procedures, the C234A *H. pylori* UPPS and the FsPP-complexed structures were refined with the addition of cofactor and solvent

Table 1: Data collection and refinement statistics for the orthorhombic *H. pylori* UPPS crystals of the apoenzyme and the complex with thiopyrophosphate. C234A mutation was included to prevent intramolecular disulfide bond formation.

| | *H. pylori* UPPS | *H. pylori* UPPS + PPi |
|---|---|---|
| | Data collection | |
| Space group | $P2_12_12_1$ | |
| Resolution (Å)[a] | 25 to 1.88 (1.95 to 1.88) | 50 to 2.5 (2.59 to 2.5) |
| Unit cell dimensions | | |
| *a*, *b*, *c* (Å) | 49.63, 58.91, 153.43 | |
| No. of reflections | | |
| Observed | 201171 (18692) | 137910 (12888) |
| Unique | 35917 (3338) | 15618 (1432) |
| Completeness (%) | 95.4 (90.3) | 96.0 (91.2) |
| $R_{\text{merge}}$ (%) | 5.5 (43.3) | 5.9 (15.9) |
| I/σ(I) | 30.7 (4.1) | 42.3 (5.2) |
| | Refinement | |
| No. of reflections[b] | 34629 (3038) | 15084 (1330) |
| $R_{\text{work}}$ (%) | 19.34 (22.91) | 21.44 (30.24) |
| $R_{\text{free}}$ (%) | 24.00 (30.02) | 29.33 (37.43) |
| Geometry deviations | | |
| Bond lengths (Å) | 0.0193 | 0.0061 |
| Bond angles (°) | 1.817 | 1.157 |
| No. of all non-H atoms | 3463 | 3449 |
| No. of water molecules | 581 | 134 |
| Mean B-values (Å$^2$) | 39.54 | 49.49 |
| Ramachandran plot (%) | | |
| Most favored | 92.1 | 92.3 |
| Additionally allowed | 7.9 | 7.7 |

[a] Values in the parentheses are for the highest resolution shells.
[b] All positive reflections are used in the refinements.

molecules. All manual modifications of the models were performed on an SGI Fuel computer using the program O [19]. Computational refinements, which included maximal likelihood and simulated-annealing protocols, were carried out using CNS. The programs MolScript [20], and Raster3D [21] were used in producing figures.

### 2.4. Computer screening to identify the inhibitors

The X-ray structure of *H. pylori* UPPS reported here and the complex structure of *E. coli* UPPS (PDB code 1V7U) were chosen as the templates in the virtual screening. The program GOLD V2.1 was used to screen Maybridge database, a commercially available compound database obtained from Maybridge Chemical Company (Tintagel, Cornwall, England). The binding pocket for the docking study was defined as a 15 Å radius sphere centered on the active site Asp13 of *H. pylori* UPPS or Asp26 of *E. coli* UPPS. The scoring function, GoldScore, implemented in GOLD was used to rank the docking positions of the compounds. 26 compounds with the highest score ranked by GoldScore were selected for inhibition assays.

### 2.5. $IC_{50}$ determination

The $IC_{50}$ values of the two hits were measured in a buffer of 100 mM Hepes (pH 7.5), 50 mM KCl, 0.5 mM $MgCl_2$, and 0.1% Triton X-100, containing 0.05 $\mu$M of *E. coli* or *H. pylori* UPPS. The concentrations of inhibitors used were ranged from 0 to 500 $\mu$M. To obtain the $IC_{50,}$ the does-response curves were fitted with the equation, $A(I) = A(0)\times\{1-[I/(I+IC_{50})]\}$, where A(I) is the enzyme activity with inhibitor concentration I, A(0) is enzyme activity without inhibitor, and I is the inhibitor concentration.

## 3. RESULTS

### 3.1. 3D structures of H. pylori UPPS

To develop structure-based inhibitors, the crystal structures of *H. pylori* UPPS were solved in this study. One is the structure of *H. pylori* UPPS containing C234A mutation to prevent intra-molecular disulfide bond formed during the long period of crystallization process (Figure 1(a)), and the other is the structure of C234A complexed with FsPP, but only the pyrophosphate portion is visible (Figure 1(b)). The C234A mutant has unchanged kinetic property compared with the wild type ($k_{\text{cat}}$, FPP $K_m$ and IPP $K_m$ of C234A were $0.20\pm0.08\ s^{-1}$, $0.15\pm0.04\ \mu$M and $9.6\pm0.2\ \mu$M, almost equal to $0.22\pm0.05\ s^{-1}$, $0.11\pm0.02\ \mu$M and $9.2\pm0.1\ \mu$M for the wild type, resp.). The overall structure of *H. pylori* UPPS was similar to that of *E. coli* UPPS [22]. The protein is a dimer and each subunit contains a catalytic domain and a pairing domain. Two subunits are tightly associated through the central $\beta$-sheet and a pair of long $\alpha$-helices ($\alpha$5 and $\alpha$6). However, *H. pylori* UPPS has a 1.5-turn shorter $\alpha$5 helix in the dimer interface. This may weaken the dimer formation for *H. pylori* UPPS. The catalytic domain is composed of six $\beta$-strands and four $\beta$-helices and the central tunnel-shaped active site is surrounded by 2 $\alpha$-helices ($\alpha$2 and $\alpha$3) and 4 $\beta$-strands ($\beta$A-$\beta$B-$\beta$D-$\beta$C) (Figure 1(a)).

At the bottom of the tunnel, a large amino acid F124 occupies a similar position to that of L137 at the bottom of *E. coli* UPPS tunnel, which is a key residue to shield the final product and determine its chain length [22]. At the top of this tunnel, several amino acids including D13, R17, R26, H30, F57, S58, R180, and E184 are located in the substrate binding site (Figure 1(b)). The position of the pyrophosphate (shown in black sticks in Figure 1(b)) of FsPP in the complex is almost identical to that of the FPP pyrophosphate in the *E. coli* UPPS active site [13].

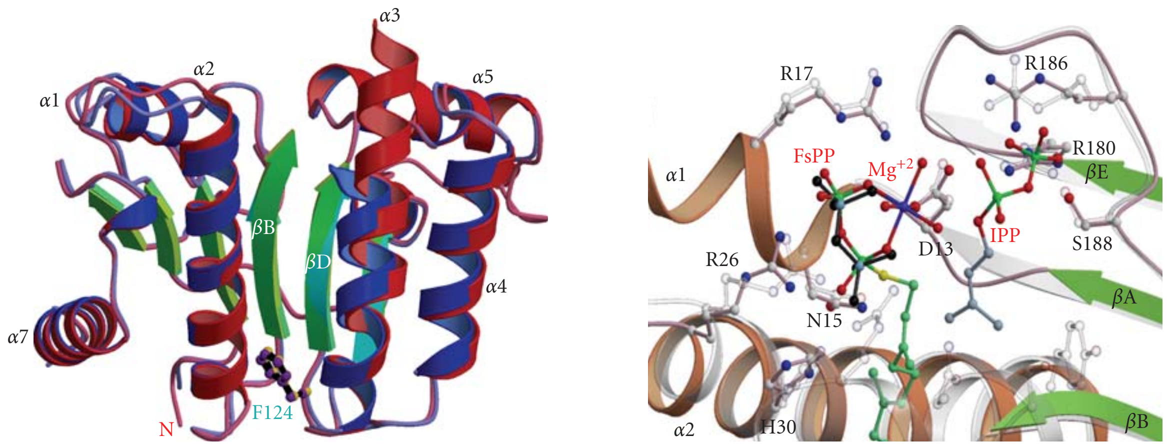
The positions of the $\alpha$3 helix in the two subunits of *H. pylori* UPPS are slightly different (Figure 1(a)), resembling the open and closed forms of *E. coli* UPPS [22]. *H. pylori* UPPS A-chain strongly resembles the Triton-bound open form of *E. coli* UPPS [23], with root mean square deviation (r.m.s.d) of 0.78 Å for 200 match pairs of $\alpha$-carbon atoms. Compare to the closed-form structure of *E. coli* UPPS with FsPP and IPP bound [9], the *H. pylori* UPPS B-chain is with the r.m.s.d. of 1.08 Å for 191 match pairs of $\alpha$-carbon atoms. This suggests a conformational change in the *H. pylori* UPPS reaction.

### 3.2. Virtual screening of the H. pylori UPPS inhibitors

Based on the structures, computer virtual screening was carried out to search for selective inhibitors of *E. coli* and *H. pylori* UPPS. The screening procedure is summarized in Figure 2. The crystal structure of *E. coli* UPPS bound with FPP (1V7U) was used as a template first for the virtual screening since the electron density of a small loop responsible for conformational change near the active site is not visible in *H. pylori* UPPS, which might confound the virtual screening result. A compound database containing 58,635 compounds available from Maybridge Chemical Company were screened using the program GOLD V2.1. Each compound in the database was docked into the active site of *E. coli* UPPS, defined as 15 Å radius sphere around Asp26, an essential residue responsible to coordinate with the catalytic $Mg^{2+}$. The docked molecules were then ranked by the GoldScore fitness function, according to the sum of H-bond energy, van der Waals energy, internal ligand van der Waals and internal torsional strain energy. The top 26 compounds ranked by GoldScore were then purchased and experimentally evaluated for their ability to inhibit *H. pylori* and *E. coli* UPPS.

### 3.3. Inhibition against E. coli and H. pylori UPPS

Of these 26 compounds, 2 compounds numbered BTB06061 and HTS04781, were found inhibitory to *H. pylori* UPPS almost equally with $IC_{50}$ values of 350 $\mu$M and 362 $\mu$M, respectively (Figure 3). The $IC_{50}$ values of these two compounds against the C234A and wild-type enzyme were almost equal. As revealed by the predicted models shown in Figures 3(a) and 3(b), two inhibitors are likely bound to *H. pylori* UPPS with a similar orientation to that of the substrate FPP. The sulfur atom in the thiazole ring of BTB06061 may form H-bonds with Asn15 and His30 while the $SO_2$ group is hydrogen bound with Met12. In addition,



(a)

(b)

Figure 1: *Crystal structures of H. pylori UPPS.* (a) Two subunits of the apoenzyme are superimposed. The most obvious disposition occurs in α3 helix which adopts an open form and a closed form in subunit A and B, respectively. At the top of the tunnel-shaped crevice surrounded by 2α-helices and 4β-strands is the substrate-binding site. Phe124 located at the bottom of the *H. pylori* UPPS tunnel adopts a similar position to that of Leu137 in *E. coli* UPPS, essential for determining product chain length. (b) Superimposition of active site structures of *H. pylori* UPPS with FsPP and *E. coli* UPPS with FsPP, $Mg^{2+}$, and IPP [9]. The active site residues in *H. pylori* UPPS are shown in pink and those in *E. coli* UPPS in white for carbon-carbon bonds in ball-and-stick model. The thiopyrophosphate (visible in crystal structure) is shown in black, the nitrogen atoms and $Mg^{2+}$ ion are shown in blue, and oxygen atoms are shown in red. Asp13 in *H. pylori* UPPS occupies a similar position to that of Asp26 in *E. coli* UPPS to coordinate with an $Mg^{2+}$ for binding with the pyrophosphate leaving group of FPP.

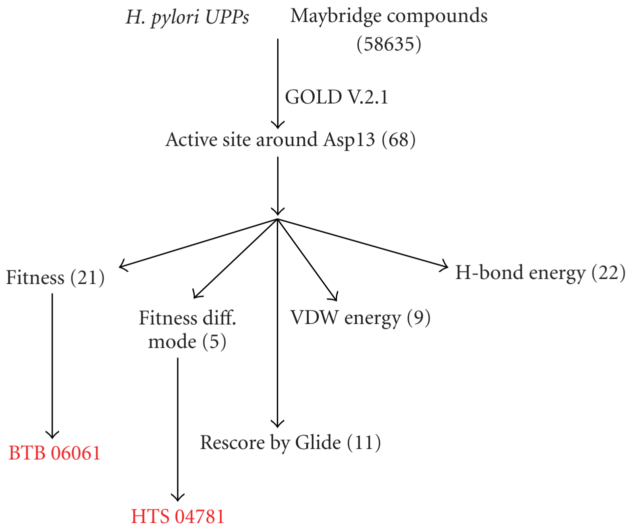


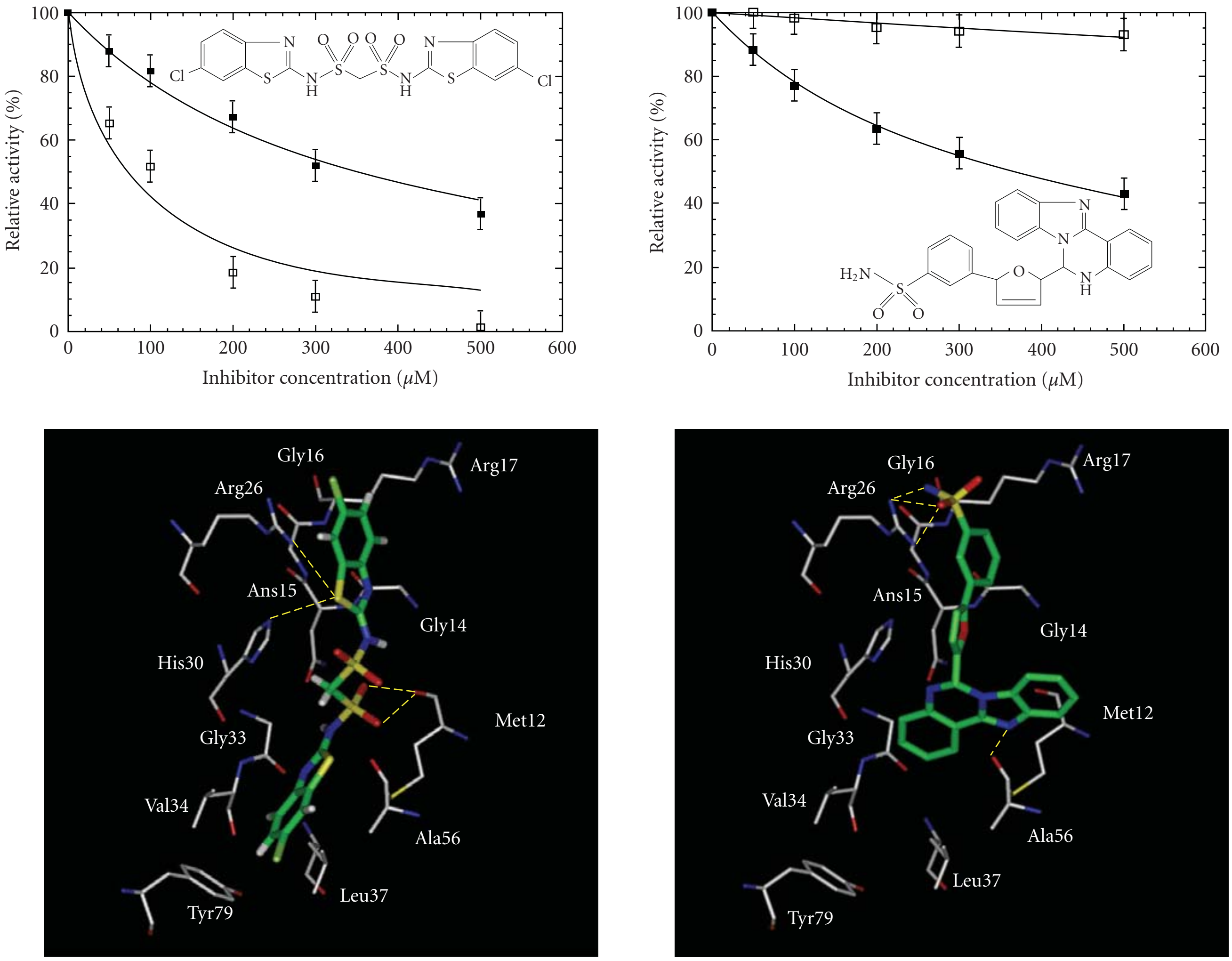
Figure 2: *The flow chart for computer screening of H. pylori UPPS inhibitors.* The active zone for screening was focused on Asp13, an important amino acid residue for coordinating with catalytic $Mg^{2+}$. In parentheses are the numbers of compounds. BTB06061 and HTS04781 are the final hits.

the aromatic rings of BTB06061 form hydrophobic interactions with the surrounding hydrophobic residues, including Val34, Leu37, Ala56 and Tyr79. As shown in the predicted model of HTS04781 with *H. pylori* UPPS, the sulfonamide group forms H-bonds with Gly16 and Arg26 and the N atom in the tetracyclic ring is hydrogen bound to the main chain of Met12. Extensive hydrophobic interactions were found between the tetracyclic ring with the surrounding residues including Met12, His30, Gly33 and Val34.

Surprisingly, BTB06061 showed 5-fold smaller $IC_{50}$ (71 μM) against *E. coli* UPPS and HTS04781 almost did not inhibit *E. coli* UPPS, although two compounds inhibited *H. pylori* equally. From the modeling (not shown), the smaller entrance in *E. coli* UPPS compared to *H. Pylori* UPPS at the top of the tunnel due to the partial blockage by the amino acids such as Trp75 from the flexible loop might restrict, or at least partially limit the access of bulky compound HTS04781 that contains four rigid aromatic rings to the active site, thereby leading to the loss of inhibitory activity when competing with the substrate for binding.

## 4. DISCUSSION

In this paper we describe the crystal structures of UPPS from *H. pyroli*, a wide-spreading and life-threatening pathogen, and the first structure-derived inhibitors from computer virtual screening. Although a high-throughput screening has been performed for UPPS by a pharmaceutical company [24], none of the inhibitors have been reported. So far, a series of IPP analogues with a dicarboxylate moiety in place of the diphosphate were synthesized and the E-pentenylbutanedioic acid showed inhibition of UPPS with an $IC_{50}$ of 135 μM [25]. Based on the known structure of UPPS (9), two carboxylate groups may coordinate with the catalytic $Mg^{2+}$ ion which was bound with the pyrophosphate group of the substrates. Recently, we reported some bisphosphonates, which inhibited *trans*-type FPPs, which could also inhibit *cis*-type UPPS with sub-μM $IC_{50}$ when containing suitable hydrophobic side-chains [26]. The crystal



(a)

(b)

Figure 3: *Computer virtual screening of the H. pylori UPPS inhibitors.* Two compounds, BTB06061 shown in (a) and HTS04781 in (b), were identified from the computer fitting of the Maybridge compounds into the active site of *E. coli* and *H. pylori* UPPS. The data of enzyme activities in the presence of different concentrations of the inhibitors were used to determine the $IC_{50}$ values of the inhibitors. The compounds displayed $IC_{50}$ of 350 and 363 μM, respectively, in inhibiting *H. pylori* UPPS activity. However, the $IC_{50}$ of BTB06061 became 71 μM in inhibiting *E. coli* UPPS and HTS04781 was almost inactive against the enzyme. The modeled structures of the inhibitor bound in the active site of *H. pylori* UPPS are shown at the bottom.

structures show that four molecules of inhibitors are bound in the active site and one of them occupies the FPP site with a phosphoate group chelating with the $Mg^{2+}$. Here, we report the first two novel inhibitors identified from a randomized compound library through virtual screening. These two inhibitors likely occupy the FPP site of *H. pylori* UPPS based on computer modeling. Two inhibitors displayed similar inhibition against *H. pylori* UPPS, but very different inhibition on *E. coli* UPPS. The one with bulky skeleton did not inhibit *E. coli* UPPS, likely owing to the partially blocked opening at the top of tunnel by the flexible loop in the *E. coli* UPPS active site. Our results shed light on the possibility of developing antibiotics specially targeting pathogenic *H. pylori* without killing the intestinal *E. coli*.

## REFERENCES

[1] C. M. Allen, “Purification and characterization of undecaprenyl pyrophosphate synthetase,” *Methods in Enzymology*, vol. 110, pp. 281–299, 1985.

[2] J. Robyt, *Essential of Carbohydrate Chemistry*, chapter 10, Springer, New York, NY, USA, 1998.

[3] K. Ogura and T. Koyama, “Enzymatic aspects of isoprenoid chain elongation,” *Chemical Reviews*, vol. 98, no. 4, pp. 1263–1276, 1998.

[4] P.-H. Liang, T.-P. Ko, and A. H.-J. Wang, “Structure, mechanism and function of prenyltransferases,” *European Journal of Biochemistry*, vol. 269, no. 14, pp. 3339–3354, 2002.

[5] A. Chen, P. A. Kroon, and C. D. Poulter, “Isoprenyl diphosphate synthases: protein sequence comparisons, phylogenetic

tree, and predictions of secondary structure," *Protein Science*, vol. 3, no. 4, pp. 600–607, 1994.

[6] N. Shimizu, T. Koyama, and K. Ogura, "Molecular cloning, expression, and purification of undecaprenyl diphosphate synthase. No sequence similarity between E- and Z-prenyl diphosphate synthases," *Journal of Biological Chemistry*, vol. 273, no. 31, pp. 19476–19481, 1998.

[7] C. M. Apfel, B. Takács, M. Fountoulakis, M. Stieger, and W. Keck, "Use of genomics to identify bacterial undecaprenyl pyrophosphate synthetase: cloning, expression, and characterization of the essential UPPS gene," *Journal of Bacteriology*, vol. 181, no. 2, pp. 483–492, 1999.

[8] H. Fujii, T. Koyama, and K. Ogura, "Efficient enzymatic hydrolysis of polyprenyl pyrophosphates," *Biochimica et Biophysica Acta*, vol. 712, no. 3, pp. 716–718, 1982.

[9] R.-T. Guo, T.-P. Ko, A. P.-C. Chen, C.-J. Kuo, A. H.-J. Wang, and P.-H. Liang, "Crystal structures of undecaprenyl pyrophosphate synthase in complex with magnesium, isopentenyl pyrophosphate, and farnesyl thiopyrophosphate: roles of the metal ion and conserved residues in catalysis," *Journal of Biological Chemistry*, vol. 280, no. 21, pp. 20762–20774, 2005.

[10] A. Nomura, G. N. Stemmermann, P. H. Chyou, G. J. Perez-Perez, and M. J. Blaser, "*Helicobacter pylori* infection and the risk for duodenal and gastric ulceration," *Annals of Internal Medicine*, vol. 120, pp. 977–981, 1994.

[11] F. Mauch, G. Bode, and P. Malfertheiner, "Identification and characterization of an ATPase system of *Helicobacter pylori* and the effect of proton pump inhibitors," *American Journal of Gastroenterology*, vol. 88, no. 10, pp. 1801–1802, 1993.

[12] P. Bytzer and C. O'Morain, "Treatment of *Helicobacter pylori*," *Helicobacter*, vol. 10, no. S1, pp. 40–46, 2005.

[13] S.-Y. Chang, T.-P. Ko, A. P.-C. Chen, A. H.-J. Wang, and P.-H. Liang, "Substrate binding mode and reaction mechanism of undecaprenyl pyrophosphate synthase deduced from crystallographic studies," *Protein Science*, vol. 13, no. 4, pp. 971–978, 2004.

[14] J.-J. Pan, L.-W. Yang, and P.-H. Liang, "Effect of site-directed mutagenesis of the conserved aspartate and glutamate on *E. coli* undecaprenyl pyrophosphate synthase catalysis," *Biochemistry*, vol. 39, no. 45, pp. 13856–13861, 2000.

[15] A. P.-C. Chen, S.-Y. Chang, Y.-C. Lin, et al., "Substrate and product specificities of cis-type undecaprenyl pyrophosphate synthase," *Biochemical Journal*, vol. 386, no. 1, pp. 169–176, 2005.

[16] Z. Otwinowski and W. Minor, "Processing of X-ray diffraction data collected in oscillation mode," *Methods in Enzymology*, vol. 276, pp. 307–326, 1997.

[17] A. T. Brunger, "Assessment of phase accuracy by cross validation: the free R value. Methods and applications," *Acta Crystallographica Section D*, vol. 49, no. 1, pp. 24–36, 1998.

[18] A. T. Brünger, P. D. Adams, G. M. Clore, et al., "Crystallography & NMR system: a new software suite for macromolecular structure determination," *Acta Crystallographica Section D*, vol. 54, no. 5, pp. 905–921, 1998.

[19] T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard, "Improved methods for building protein models in electron density maps and the location of errors in these models," *Acta Crystallographica Section A*, vol. 47, no. 2, pp. 110–119, 1991.

[20] P. J. Kraulis, "*MOLSCRIPT*: a program to produce both detailed and schematic plots of protein structures," *Journal of Applied Crystallography*, vol. 24, no. 5, pp. 947–950, 1991.

[21] E. A. Merritt and M. E. P. Murphy, "Raster3D version 2.0 A program for photorealistic molecular graphics," *Acta Crystallographica Section D*, vol. 50, no. 6, pp. 869–873, 1994.

[22] T.-P. Ko, Y.-K. Chen, H. Robinson, et al., "Mechanism of product chain length determination and the role of a flexible loop in *Escherichia coli* undecaprenyl-pyrophosphate synthase catalysis," *Journal of Biological Chemistry*, vol. 276, no. 50, pp. 47474–47482, 2001.

[23] S.-Y. Chang, T.-P. Ko, P.-H. Liang, and A. H.-J. Wang, "Catalytic mechanism revealed by the crystal structure of undecaprenyl pyrophosphate synthase in complex with sulfate, magnesium, and triton," *Journal of Biological Chemistry*, vol. 278, no. 31, pp. 29298–29307, 2003.

[24] H. Li, J. Huang, X. Jiang, M. Seefeld, M. McQueney, and R. Macarron, "The effect of triton concentration on the activity of undecaprenyl pyrophosphate synthase inhibitors," *Journal of Biomolecular Screening*, vol. 8, no. 6, pp. 712–715, 2003.

[25] A. A. Scholte, L. M. Eubanks, C. D. Poulter, and J. C. Vederas, "Synthesis and biological activity of isopentenyl diphosphate analogues," *Bioorganic and Medicinal Chemistry*, vol. 12, no. 4, pp. 763–770, 2004.

[26] R.-T. Guo, R. Cao, P.-H. Liang, et al., "Bisphosphonates target multiple sites in both cis- and trans- prenyltransferases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 24, pp. 10022–10027, 2007.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 768232, 7 pages
doi:10.1155/2008/768232

*Research Article*

# A Robotic Voice Simulator and the Interactive Training for Hearing-Impaired People

**Hideyuki Sawada, Mitsuki Kitani, and Yasumori Hayashi**

*Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University, Japan*

Correspondence should be addressed to Hideyuki Sawada, sawada@eng.kagawa-u.ac.jp

Received 31 August 2007; Accepted 23 January 2008

Recommended by Daniel Howard

A talking and singing robot which adaptively learns the vocalization skill by means of an auditory feedback learning algorithm is being developed. The robot consists of motor-controlled vocal organs such as vocal cords, a vocal tract and a nasal cavity to generate a natural voice imitating a human vocalization. In this study, the robot is applied to the training system of speech articulation for the hearing-impaired, because the robot is able to reproduce their vocalization and to teach them how it is to be improved to generate clear speech. The paper briefly introduces the mechanical construction of the robot and how it autonomously acquires the vocalization skill in the auditory feedback learning by listening to human speech. Then the training system is described, together with the evaluation of the speech training by auditory impaired people.

## 1. INTRODUCTION

A voice is the most important and effective medium employed not only in daily communication but also in logical discussions. Only humans are able to use words as means of verbal communication, although almost all animals have voices. Vocal sounds are generated by the relevant operations of the vocal organs such as a lung, trachea, vocal cords, vocal tract, tongue, and muscles. The airflow from the lung causes a vocal cord vibration to generate a source sound, and then the glottal wave is led to the vocal tract, which works as a sound filter as to form the spectrum envelope of a particular voice. The voice is at the same time transmitted to the auditory system so that the vocal system is controlled for the stable vocalization. Different vocal sounds are generated by the complex movements of vocal organs under the feedback control mechanisms using an auditory system.

As infants grow they acquire these control methods pertaining to the vocal organs for appropriate vocalization. These get developed in infancy by repetition of trials and errors concerning the hearing and vocalizing of vocal sounds. Any disability or injury to any part of the vocal organs or to the auditory system may result in an impediment in vocalization. People who have congenitally hearing impairments have difficulties in learning vocalization, since they are not able to listen to their own voice. A speech therapist helps themtotrain their speech by teaching the vocal organs to learn vocalization and clear speech [1–4].

We are developing a talking robot by reproducing a human vocal system mechanically and based on the physical model of the vocal organs in the human. The fundamental frequency and the spectrum envelope determine the principal characteristics of a voice. Fundamental frequency is a characteristic of the voice source that is generated by the vibration of vocal cords. The resonance effects that get articulated by the motion of vocal tract and nasal cavity cause the spectrum envelope. For the autonomous acquisition of vocalization skills by the robot, an adaptive learning using an auditory feedback control is introduced, like the case for a human baby.

The robot consists of motor-controlled vocal organs such as vocal cords, a vocal tract, and a nasal cavity to generate a natural voice imitating a human vocalization [5–8]. By introducing auditory feedback learning with an adaptive control algorithm of pitch and phoneme, the robot is able to autonomously acquire the control skill of the mechanical system to vocalize stable vocal sounds imitating human speech. In the first part of the paper, the construction of vocal cords

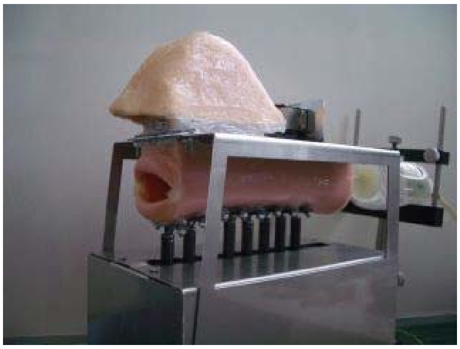


Figure 1: Structural view of talking robot.

and vocal tract for the realization of the robot is briefly presented, and then the analysis of the autonomous learning of how the robot acquires the vocalization skill by using the neural network will be described. Then, a robotic training system for the hearing-impaired people is introduced, together with the evaluation of the interactive speech training conducted in an experiment.

## 2. CONSTRUCTION OF A TALKING ROBOT

The talking robot mainly consists of an air pump, artificial vocal cords, a resonance tube, a nasal cavity, and a microphone connected to a sound analyzer, which, respectively, correspond to a lung, vocal cords, a vocal tract, a nasal cavity, and an audition of a human, as shown in Figure 1.

An air from the pump is led to the vocal cords via an airflow control valve, which works for the control of the voice volume. The resonance tube as a vocal tract is attached to the vocal cords for the modification of resonance characteristics. The nasal cavity is connected to the resonance tube with a sliding valve between them. The sound analyzer plays a role of the auditory system. It realizes the pitch extraction and the analysis of resonance characteristics of generated sounds in real time, which are necessary for the auditory feedback control. The system controller manages the whole system by listening to the vocalized sounds and calculating motor control commands, based on the auditory feedback control mechanism employing a neural network learning. The relation between the voice characteristics and motor control parameters is stored in the system controller, which is referred to in the generation of speech and singing performance.

### *2.1. Artificial vocal cords and its pitch control*

Vocal cords with two vibrating cords molded with silicone rubber with the softness of human mucous membrane were constructed in this study. Two-layered construction (a hard silicone is inside with the soft coating outside) gave the better resonance characteristics, and is employed in the robot [7]. The vibratory actions of the two cords are excited by the airflow led by the tube, and generate a source sound to be resonated in the vocal tract.

The tension of cords can be manipulated by applying tensile force to them. By pulling the cords, the tension increases so that the frequency of the generated sound becomes higher. The relationship between the tensile force and the fundamental frequency of a vocal sound generated by the robot is acquired by the auditory feedback learning before the singing and talking performance, and pitches during the utterance are kept in stable by the adaptive feedback control [8].

### *2.2. Construction of resonance tube and nasal cavity*

The human vocal tract is a non-uniform tube about 170 mm long in man. Its cross-sectional area varies from 0 to 20 cm$^2$ under the control for vocalization. A nasal cavity with a total volume of 60 cm$^3$ is coupled to the vocal tract. In the mechanical system, a resonance tube as a vocal tract is attached at the sound outlet of the artificial vocal cords. It works as a resonator of a source sound generated by the vocal cords. It is made of a silicone rubber with the length of 180 mm and the diameter of 36 mm, which is equal to 10.2 cm$^2$ by the cross-sectional area as shown in Figure 1. The silicone rubber is molded with the softness of human skin, which contributes to the quality of the resonance characteristics.

In addition, a nasal cavity made of a plaster is attached to the resonance tube to vocalize nasal sounds like /*m*/ and /*n*/. A sliding valve as a role of the soft palate is settled at the connection of the resonance tube and the nasal cavity for the selection of nasal and normal sounds. For the generation of nasal sounds /*n*/ and /*m*/, the motor-controlled sliding valve is open to lead the air into the nasal cavity.

By actuating displacement forces with stainless bars from the outside of the vocal tract, the cross-sectional area of the tube is manipulated so that the resonance characteristics are changed according to the transformations of the inner areas of the resonator. Compact servo motors are placed at 8 positions $x_j$ $(j = 1–8)$ from the lip side of the tube to the intake side, and the displacement forces $P_j(x_j)$ are applied according to the control commands from the motor-phoneme controller.

## 3. LEARNING OF VOCALIZATION SKILL

An adaptive learning algorithm for the achievement of a talking and singing performance is introduced in this section. The algorithm consists of two phases. First in the learning phase, the system acquires two maps in which the relations between the motor positions and the features of generated voices are established and stored. One is a motor-pitch map, which associates motor positions with fundamental frequencies. It is acquired by comparing the pitches of vocalized sounds with the desired pitches, which cover the frequency range of speech [8]. The other is a motor-phoneme map, which associates motor positions with phonetic features of vowel and consonant sounds. Second in the performance phase, the robot speaks and sings by referring to the obtained maps, while pitches and phonemes of generated voices are adaptively maintained by hearing its own output voices.

### 3.1. Neural network learning of vocalization

The neural network (NN) works to associate the sound characteristics with the control parameters of the nine motors settled in the vocal tract and the nasal cavity. In the learning process, the network learns the motor control commands by inputting 10th-order linear predictive coding (LPC) cepstrum coefficients [9] derived from vocal sound waves as teaching signals. The network acquires the relations between the sound parameters and the motor control commands of the vocal tract. After the learning, the neural network is connected in series into the vocal tract model. By inputting the sound parameters of desired sounds to the NN, the corresponding form of the vocal tract is obtained.

In this study, the self-organizing neural network (SONN) was employed for the adaptive learning of vocalization. Figure 2 shows the structure of the SONN consisting of two processes, which are an information memory process and an information recall process. After the SONN learning, the motor control parameters are adaptively recalled by the stimuli of sounds to be generated.

The information memory process is achieved by the self-organizing map (SOM) learning [10], in which sound parameters are arranged onto a two-dimensional feature map to be related to one another.

Weight vector $\mathbf{V}_j$ at node $j$ in the feature map is fully connected to the input nodes $x_i$ $[i = 1, \dots, 10]$, where 10th-order LPC cepstrum coefficients are given. The map learning algorithm updates the weight vectors $\mathbf{V}_j$-s. A competitive learning is used, in which the winner $c$ as the output unit with a weight vector closest to the current input vector $\mathbf{x}(t)$ is chosen at time $t$ in learning. By using the winner $c$, the weight vectors $\mathbf{V}_j$-s are updated according to the rule shown below;

$$\mathbf{V}_j(t+1) = \mathbf{V}_j(t) + h_{cj}(t)\left[\mathbf{x}(t) - \mathbf{V}_j(t)\right],$$
$$h_{cj}(t) = \begin{cases} \alpha(t)\cdot\exp\left(-\dfrac{\left\|r_c - r_j\right\|^2}{2\sigma^2(t)}\right) & (i \in N_c), \\ 0 & (i \notin N_c). \end{cases} \tag{1}$$

Here, $\|r_c - r_j\|$ is the distance between units $c$ and $j$ in the output array, and $N_c$ is the neighborhood of the node $c$. $\alpha(t)$ is a learning coefficient which gradually reduces as the learning proceeds. $\sigma(t)$ is also a coefficient which represents the width of the neighborhood area.

Then, in the information recall process, each node in the feature map is associated with motor control parameters for the control commands of nine motors employed for the vocal tract deformation, by using the three-layered perceptron. In this study, a conventional back-propagation algorithm was employed for the learning. With the integration of the information memory and recall processes, the SONN works to adaptively associate sound parameters with motor control parameters.

In the current system, 25 × 25 arrayed map $\mathbf{V} = [V_1, V_2, \dots, V_{25\times 25}]$ is used as the SOM. For testing the mapping ability, 200 sounds randomly vocalized by the robot were mapped onto the map array. After the self-organizing learning, five Japanese vowels vocalized by six different people were mapped onto the feature map. Same vowel sounds given by different people were mapped close with each other, and five vowels were roughly categorized according to the differences of phonetic characteristics. We found that, in some vowel area, two sounds given by two different speakers fell in a same unit in the feature map. It means that the two different sounds could not be separated, although they have close tonal features with each other. We propose a reinforcement learning algorithm to optimize the feature map.

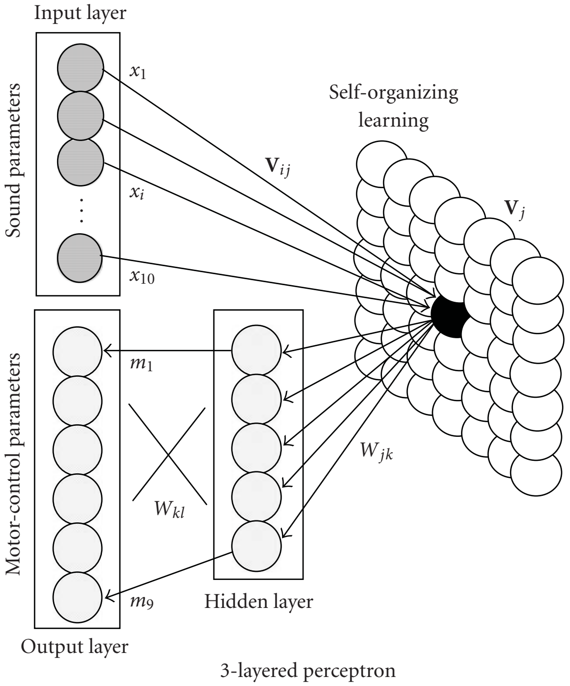


Figure 2: Structure of self-organizing neural network.

### 3.2. Reinforcement learning of five Japanese vowels by human voices

Redundant sound parameters which were not used for the Japanese speech were buried in the map, since the 150 inputted sounds were generated randomly by the robot. Furthermore, two different sounds given by two different speakers were occasionally fallen in the same unit. The mapping should be optimized for the Japanese vocalization.

The reinforcement learning was employed to establish the feature map optimized. After the SONN learning, five Japanese vowel sounds given by 6 different speakers with normal audition were applied to the supervised learning as the reinforcement signal to be associated with the suitable motor control parameters for the Japanese vocalization.

Figure 3 shows the result of the reinforcement learning with five Japanese vowels given by five speakers no. 1 to 5. The distribution of same vowel sounds concentrated with one another, and the patterns of different vowels were placed apart.

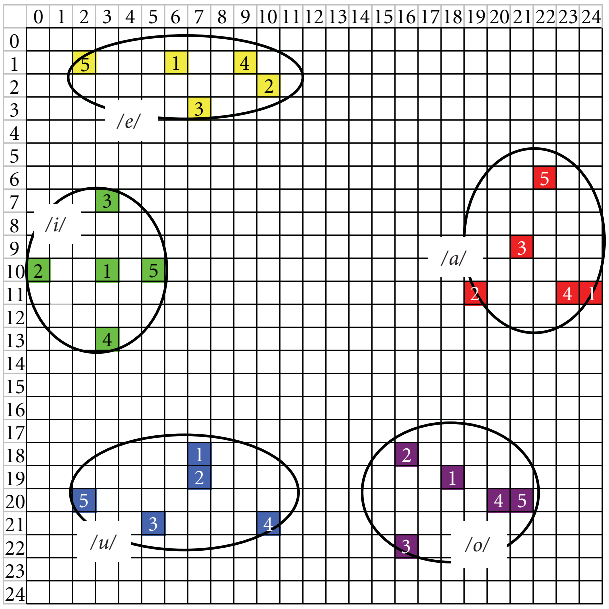


FIGURE 3: Result of reinforcement learning with five Japanese vowels from 5 subjects no. 1–5.

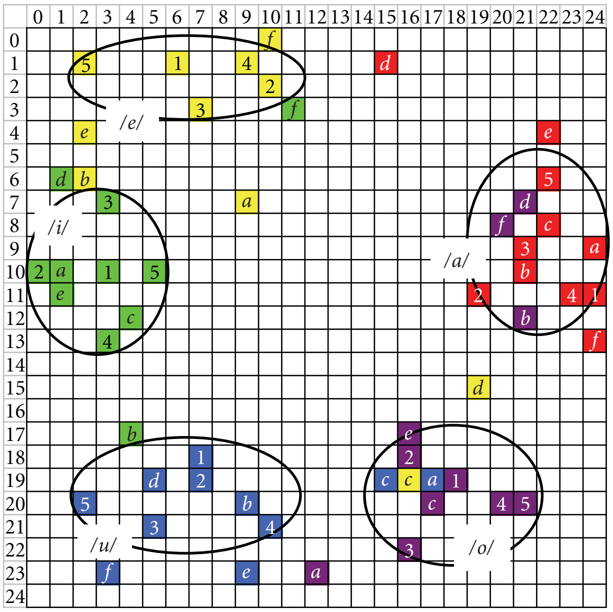


FIGURE 4: Mapping results of six different voices given by hearing-impaired speakers no. a–c.

## 4. ARTICULATORY REPRODUCTION OF HEARING-IMPAIRED VOICE

After the learning of the relationship between the sound parameters and the motor control parameters, we inputted human voices from microphone to confirm whether the robot could speak autonomously by mimicking human vocalization. With the comparison of spectra between human vowel vocalization and robot speech, we confirmed that the first and second formants F1 and F2, which present the principal characteristics of the vowels, were formed properly as to approximate the human vowels, and the sounds were well distinguishable by listeners. The experiment also showed the smooth motion of the vocalization. The transition between two different vowels in the continuous speech was well acquired by the SONN learning, which means that all the cellsoninthe SOM are associated with motor control parameters properly to vocalize particular sounds [11].

Voices of hearing-impaired people then were given to the robot so as to confirm that the articulatory motion would be reproduced by the robot. Figure 4 shows the mapping results of six different voices given by hearing-impaired speakers no. a, no. b, no. c, no. d, no. e, and no. f. The same colors indicate the vocal sounds generated by the same vowels. In Figure 5, vocal tract shapes estimated by the robot from voices of hearing-impaired person no. a are presented, together with the comparison of the vocal tract shapes estimated by the able-bodied speaker no. 1 voices.

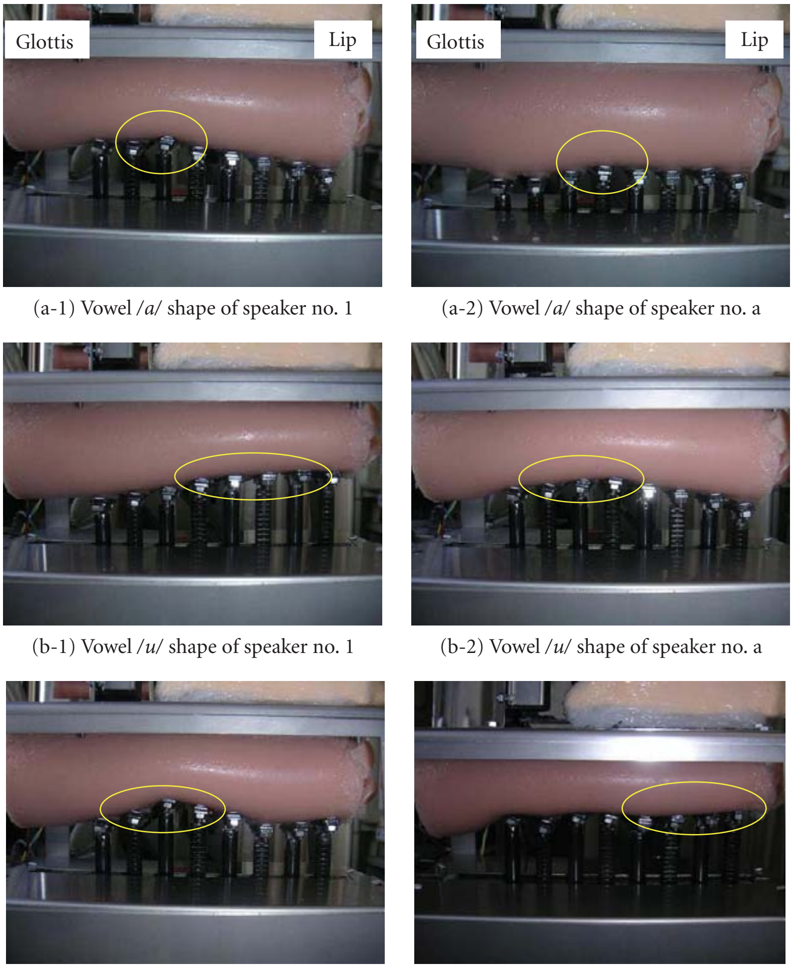
From the observation of the robot's reproduced motions of the vocal tract, the articulations of auditory-impaired people were apparently small, and complex shapes of vocal tract were not sufficiently articulated. Furthermore, in the map shown in Figure 4, /*u*/ sound given by the hearing-impaired speaker no. a is located inside the /*e*/ area of able-bodied speakers, and his /*o*/ vowel is located close to the /*u*/ area of able-bodied speakers. These articulatory characteristics also appear in the vocal tract shapes shown in Figure 5. In the figures, the vowel /*u*/ shape of speaker no. a shown in (b-2) is almost the same with the /*o*/ shape of speaker no. 1 presented in (c-1). Likewise, the /*o*/ shape shown in (c-2) appears close to the shape of (b-1). Thus, these results proved that the topological relations of resonance characteristics of voices were well preserved in the map, and the articulatory motion by the robot was successfully obtained to reproduce the speech articulation by listening arbitrary vocal sounds.

## 5. INTERACTIVE VOICE TRAINING SYSTEM FOR HEARING-IMPAIRED PEOPLE

In the speech training, the robot interactively shows the articulatory motion of vocal organs as a target to a trainee so thats/he repeats his/her vocalization and the observation of the robot motion. The trainee is also able to refer to the SOM to find the distance to the target voice. The flow of the training is summarized in Figure 6. The training of speech articulation by an auditory-impaired subject is shown in Figure 7.

### *Subject*

An experiment of speech training was conducted by six hearing-impaired subjects: no. a–f (four males and two females), who study in a high school and a junior high school. In Figure 8, the training results of three subjects no. a, no. e, and no. f are shown by presenting the trajectories of voices appeared in the SOM during the training experiments. Figure 8(a) shows a result of successful training with less trials conducted by the subject no. a. By observing the articulatory motion instructed by the robot, this subject recognized



(a-1) Vowel /*a*/ shape of speaker no. 1

(a-2) Vowel /*a*/ shape of speaker no. a

(b-1) Vowel /*u*/ shape of speaker no. 1

(b-2) Vowel /*u*/ shape of speaker no. a

(c-1) Vowel /*o*/ shape of speaker no. 1

(c-2) Vowel /*o*/ shape of speaker no. a

Figure 5: Comparison of vocal tract shapes of the hearing-impaired (right) with the able-bodied (left).

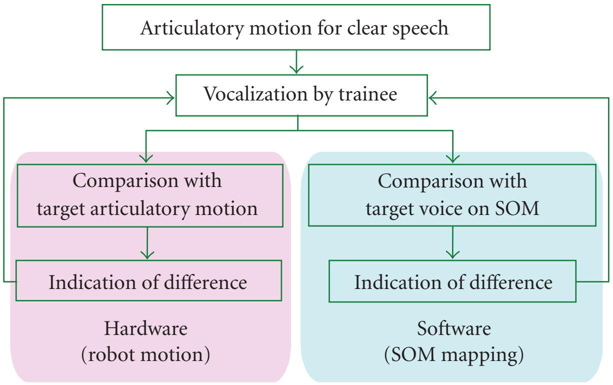


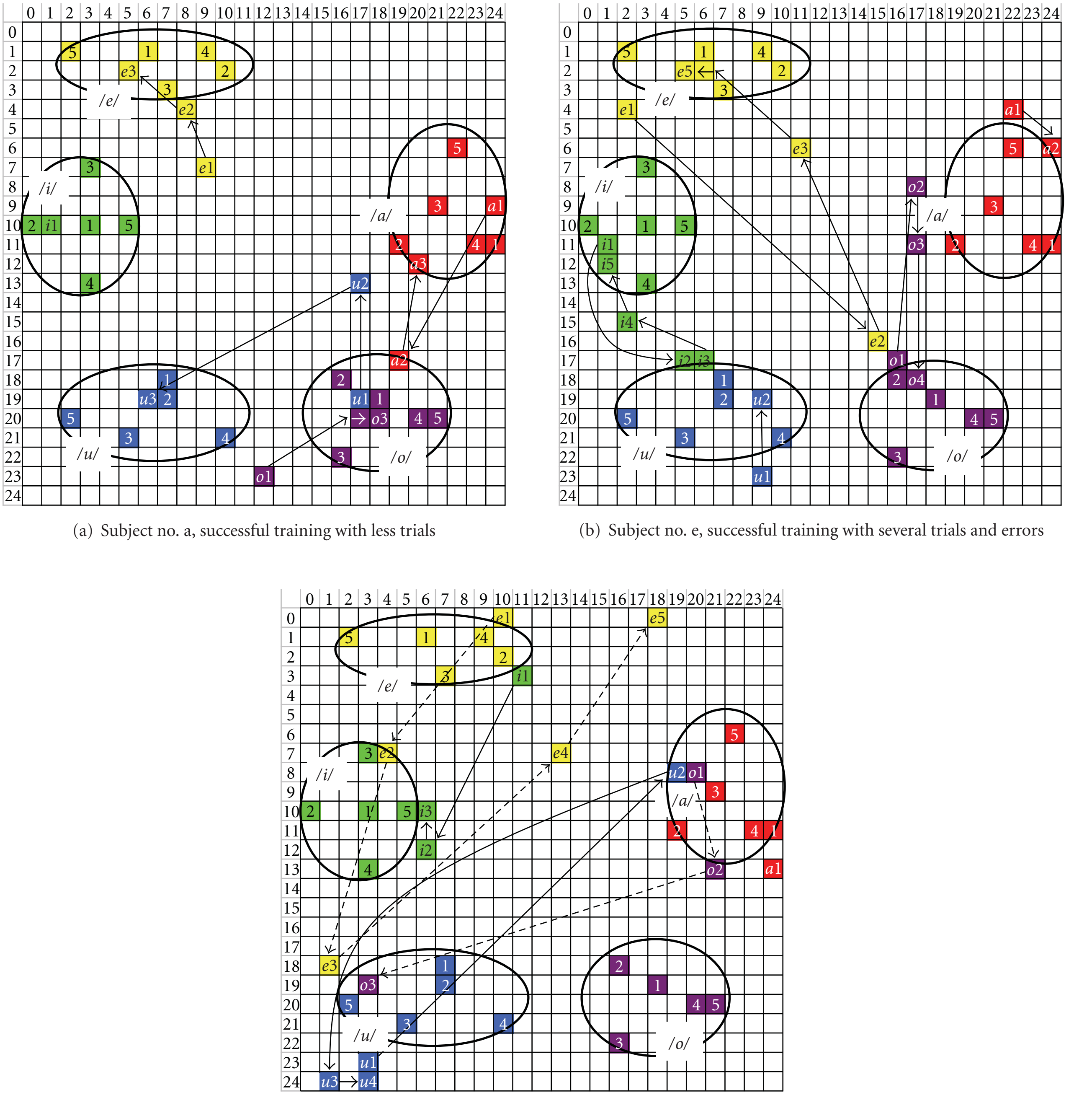
Figure 6: Flowchart of training of speech articulation.

[redacted]

Figure 7: Training of speech articulation by auditory-impaired people.

the difference in his articulation and effectively learned the correct motion. Figure 8(b) also shows the successful training results by the subject no. e, however, he had achieved the vocalization by repeating several trials and errors, especially for the vowels /*i*/ and /*e*/ as presented by the arrows from *i1* to *i5* and *e1* to *e5*, respectively.

In the case of the training conducted by the subject no. f, he could not achieve the learning by the system. The clarity of



(a) Subject no. a, successful training with less trials

(b) Subject no. e, successful training with several trials and errors

(c) Subject no. f, fail of training

Figure 8: Example trajectories in training.

his voices was quite low, and the original voices were mapped far from the area of clear voices. He could not understand the shape of the robot's vocal tract, nor realize the correspondence between the robot's motion and the motion of his inner mouth. This subject tried to articulate his vocal tract following the articulatory motion indicated by the robot, however, his voice moved to the different direction in the SOM as shown by arrows in Figure 8(c). He failed the acquisition of vocalization skill and could not achieve the training. In the questionnaire after the training, he pointed out the difficulties of moving a particular part of the inner mouth so as to mimic the articulatory motion of the robot.

By the experimental training, five subjects could mimic the vocalization following the directions given by the robotic voice simulator, and acquired the better vocal sounds. In the questionnaire after the experiment, two subjects commented that the correspondence between robot's vocal tract and human actual vocal tract should be instructed, so that they could easily understand which part inside the mouth should be intensively articulated for the clear vocalization.

## 6. CONCLUSIONS

A robotic voice simulator and its articulatory reproduction of voice of hearing-impaired people were introduced in this paper. By introducing the adaptive learning and controlling of the mechanical model with the auditory feedback, the voice robot was able to acquire the vocalization skill as a human baby does in speech training.

The robot was applied to introduce a training system for auditory-impaired people to interactively train the speech articulation for learning proper vocalization. The robotic voice simulator reproduces the articulatory motion just by listening to actual voices given by auditory-impaired people, and they could learn and know how to move their vocal organs for the clear vocalization, by observing the motions instructed by the talking robot. The use of SOM for visually presenting the distance between target voice and trainee's voice is also introduced.

We confirmed that the training using the talking robot and the SOM would help hearing-impaired people learn the articulatory motion in the mouth and the skill of clear vocalization properly. In the next system, the correspondence between robot's vocal tract and human actual vocal tract should be established so that a subject could understand which part inside the mouth should be intensively articulated in the training. By analyzing the vocal articulation of auditory-impaired people during the training with the robot, we will investigate the factor of unclarity of their voices originated by the articulatory motions.

## ACKNOWLEDGMENTS

This work was partly supported by the Grants-in-Aid for Scientific Research, the Japan Society for the Promotion of Science (no. 18500152). The authors would like to thank Dr. Yoichi Nakatsuka, the director of the Kagawa Prefectural Rehabilitation center for the Physically Handicapped, Mr. Tomoyoshi Noda, the speech therapist and teacher of Kagawa Prefectural School for the Deaf, and the students of the school for their helpful supports for the experiment and the useful advice.

## REFERENCES

[1] A. Boothroyd, *Hearing Impairments in Young Children*, Alexander Graham Bell Association for the Deaf, Washington, DC, USA, 1988.

[2] A. Boothroyd, "Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display," *IEEE Transactions on Audio and Electroacoustic*, vol. 21, no. 3, pp. 274–278, 1973.

[3] N. P. Erber and C. L. de Filippo, "Voice/mouth synthesis and tactual/visual perception of /pa, ba, ma/," *Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1015–1019, 1978.

[4] M. H. Goldstein and R. E. Stark, "Modification of vocalizations of preschool deaf children by vibrotactile and visual displays," *Journal of the Acoustical Society of America*, vol. 59, no. 6, pp. 1477–1481, 1976.

[5] H. Sawada and S. Hashimoto, "Adaptive control of a vocal chord and vocal tract for computerized mechanical singing instruments," in *Proceedings of the International Computer Music Conference (ICMC '96)*, pp. 444–447, Hong Kong, September 1996.

[6] T. Higashimoto and H. Sawada, "Vocalization control of a mechanical vocal system under the auditory feedback," *Journal of Robotics and Mechatronics*, vol. 14, no. 5, pp. 453–461, 2002.

[7] T. Higashimoto and H. Sawada, "A mechanical voice system: construction of vocal cords and its pitch control," in *Proceeding of the 4th International Conference on Intelligent Technologies (InTech '03)*, pp. 762–768, Chiang Mai, Thailand, December 2003.

[8] H. Sawada, M. Nakamura, and T. Higashimoto, "Mechanical voice system and its singing performance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, vol. 2, pp. 1920–1925, Sendai, Japan, September-October 2004.

[9] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 1995.

[10] J. D. Markel, *Linear Prediction of Speech*, Springer, New York, NY, USA, 1976.

[11] M. Nakamura and H. Sawada, "Talking robot and the analysis of autonomous voice acquisition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 4684–4689, Beijing, China, October 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 472719, 7 pages
doi:10.1155/2008/472719

*Research Article*

# Unsupervised Learning in Detection of Gene Transfer

**L. Hamel,[1] N. Nahar,[1] M.S. Poptsova,[2] O. Zhaxybayeva,[3] and J.P. Gogarten[2]**

[1] *Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881, USA*
[2] *Department of Molecular and Cell Biology, College of Liberal Arts and Sciences, University of Connecticut, CT 06269, USA*
[3] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 1X5, Canada*

Correspondence should be addressed to L. Hamel, hamel@cs.uri.edu

Received 13 September 2007; Accepted 14 December 2007

Recommended by Daniel Howard

The tree representation as a model for organismal evolution has been in use since before Darwin. However, with the recent unprecedented access to biomolecular data, it has been discovered that, especially in the microbial world, individual genes making up the genome of an organism give rise to different and sometimes conflicting evolutionary tree topologies. This discovery calls into question the notion of a single evolutionary tree for an organism and gives rise to the notion of an evolutionary consensus tree based on the evolutionary patterns of the majority of genes in a genome embedded in a network of gene histories. Here, we discuss an approach to the analysis of genomic data of multiple genomes using bipartition spectral analysis and unsupervised learning. An interesting observation is that genes within genomes that have evolutionary tree topologies, which are in substantial conflict with the evolutionary consensus tree of an organism, point to possible horizontal gene transfer events which often delineate significant evolutionary events.

## 1. INTRODUCTION

Evolutionary history of species is now inferred from the evolutionary histories of their genomes. Genomes can be viewed as a collection of genes and whole genome evolution is concluded from the evolution of individual genes. If the majority of genes followed the same evolutionary history, supertree approaches can be used to calculate a majority consensus tree. However, evolutionary trees of individual genes can differ from the majority [1], and in this case, the consensus tree is embedded in a network represented by the histories of the different genes. Evolutionary tree topologies of genes that conflict with the consensus tree are strong indicators of horizontal gene transfer events. Given this, it is clear that organismal evolution cannot be inferred from studying the evolution of just a few genes but must be inferred from studying as many (orthologous) genes as possible.

To construct and evaluate an evolutionary consensus tree based on multiple genes for a set of genomes, it is advisable to construct all possible evolutionary tree topologies for these genomes and measure the support of each topology by the (orthologous) genes within the genomes. Unfortunately, evaluating all possible tree topologies is computationally intractable for any but a very small set of genomes, since the number of possible tree topologies grows factorially with the number of participating genomes. An approach based on the spectral analysis of genomic data using bipartitions [2, 3] allows the inference of consensus trees from smaller quanta of phylogenetic information, side stepping some of the difficult computational issues. Table 1 shows the number of possible trees versus the number of possible bipartitions given a fixed set of genomes. With $n$ taxa there are $(2n-5)!/[2^{(n-3)}(n-3)!]$ different unrooted tree topologies. The number of possible nontrivial bipartitions for $n$ taxa is given by the formula $2^{(n-1)} - n - 1$, and it grows much slower with an increasing number of species than the number of different trees. We refer to the approach based on bipartitions as *spectral genome analysis*.

It is worth noting that when a single tree is calculated from the combination of all genes, including genes that were horizontally transferred, the topology of the resulting tree might not represent the plurality of gene histories. Therefore, a detailed analysis of the evolutionary histories of the participating genes is of interest. The techniques outlined here support this kind of analysis.

In spectral genome analysis, each set of orthologous genes (a gene family) is associated with a particular set of bipartitions (its *spectrum*) that define its evolutionary tree.

Table 1: Number of possible trees and bipartitions given a fixed set of genomes.

| Number of genomes | Number of unrooted trees | Number of nontrivial bipartitions |
|---|---|---|
| 4 | 3 | 3 |
| 5 | 15 | 10 |
| 6 | 105 | 25 |
| 7 | 945 | 56 |
| 8 | 10,395 | 119 |
| 9 | 135,135 | 246 |
| 10 | 2,075,025 | 501 |
| 20 | 2.22E + 20 | 5.24E + 05 |
| 50 | 2.84E + 74 | 5.63E + 14 |
| $n$ | $(2n-5)!/[2^{(n-3)}(n-3)!]$ | $2^{(n-1)} - n - 1$ |

Thus, we can envision a gene family as a point in the space spanned by all possible bipartitions of a set of genomes. Here, we apply unsupervised learning in the form of self-organizing maps [4] to this space and obtain a visual representation of clusters of gene families with similar spectra. The spectra of the gene families within a particular cluster allow us to infer the consensus tree for that cluster. It is now possible to investigate whether the consensus tree topologies of the clusters are compatible or conflicting with the overall consensus tree. If a cluster of gene families is discovered that conflicts with the consensus tree topology, then this is a strong indication for a horizontal gene transfer event. The advantage of this approach is that we not only see a distinction between consensus and conflicting trees, but that we can detect trends of agreement between the conflicting genes. This additional insight might provide biological clues as to the nature of the origin of these genes.

Unsupervised learning has been used in genomic analyses before (e.g., [5]). However, our approach seems to be novel in that we do not apply unsupervised learning directly to DNA sequence data but instead analyze the much more abstract representation of the genomic data in the form of bipartitions. We have constructed a web service called Gene Phylogeny eXplorer (GPX, http://bioinformatics.cs.uri.edu/gpx)) that supports spectral genome analysis [6].

## 2. MATERIALS AND METHODS

### 2.1. Spectral analysis of evolutionary trees

Given $n$ entities, there are $2^{n-1} - 1$ different ways to assign the entities to two different nonempty sets. That is, there are $2^{n-1} - 1$ different *bipartitions* of $n$ entities including trivial bipartitions. An (unrooted) tree can be viewed as a model of the evolutionary relationships between $n$ entities or taxa such as species, genes, molecules, and so forth. Trees and bipartitions are related as follows. Each edge in a tree can be seen as dividing the tree into a bipartition: the leaf nodes that can be reached from one end of the edge form one set of taxa and the leaf nodes that can be reached from the other end of the edge form the other set of taxa. A binary tree with $n$ leaf nodes has exactly $2n - 3$ edges. Thus, an evolutionary tree relating $n$ taxa gives rise to $2n - 3$ bipartitions. It is easy to see that $2n - 3 < 2^{n-1} - 1$, that is, the number of bipartitions defined by an evolutionary binary tree of $n$ taxa is much smaller than the number of possible bipartitions of $n$ entities.

Trivial bipartitions, which is bipartitions where one of the partitions is a singleton set, do not contain any phylogenetic information. Thus, given $n$ entities, there are $2^{(n-1)} - n - 1$ different nontrivial bipartitions. However, in an unrooted binary tree with $n$ leaf nodes there are $n - 3$ interior edges and therefore $n - 3$ nontrivial bipartitions. An interior edge is an edge that is not incident to a leaf node of a tree. It is evident that $n - 3 < 2^{n-1} - n - 1$, that is, the number of nontrivial bipartitions generated by a tree is much smaller than the number of possible nontrivial bipartitions.

Let $t_n$ be an evolutionary tree over $n$ taxa, then we define the bipartitions of $t_n$ as the *spectrum* of $t_n$, denoted as $S(t_n)$. It is convenient to adopt a vector notation for the spectrum $S(t_n) = (b_1, \ldots, b^n_{2-1}) = (0, 1, 1, 0, \ldots, 0, 0)$, where $b_k$ denotes bipartition $k$ with $1 < k < 2^{n-1} - 1$. Here, $b_k = 1$ if the spectrum of the tree includes bipartition $b_k$, and $b_k = 0$ otherwise. Note that the vector notation is a representation over all possible bipartitions. Given this, we can now refer to a *bipartition space* and we can readily see that a spectrum of a particular evolutionary tree $t_n$ represents the coordinates of a point in that space. In our case, where the tree represents the evolutionary relationship between orthologous genes in $n$ genomes, we often refer to the spectrum as the gene family spectrum and therefore a gene family is denoted by a point in bipartition space.

Figure 1(a) is an unrooted tree relating five taxa A through E. The arrows indicate branches defining the nontrivial bipartitions in this tree. Figure 1(b) represents a bipartition corresponding to the left arrow in Figures 1(a) and 1(c) represents a bipartition corresponding to the right arrow in Figure 1(a), respectively. Observe that the sub-tree topologies in the bipartitions are unresolved.

By further generalizing and interpreting the values in the spectrum vectors as arbitrary real numbers, as we will do in what follows when we assign confidence values to bipartitions, a bipartition space can be viewed as a $2^{n-1} - 1$ dimensional real vector space. An interesting consequence of this is that we can now measure the difference between spectra as the Euclidean distance between the twocorresponding spectrum points in a bipartition space.

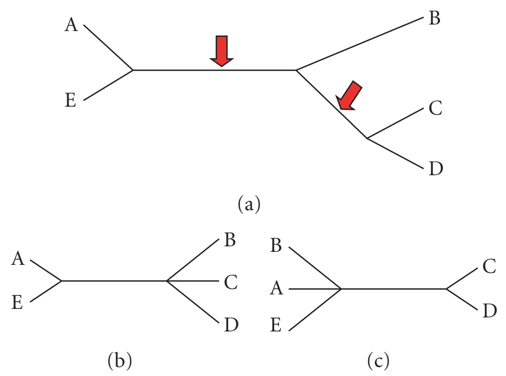


Figure 1: (a) An unrooted tree with 5 taxa, (b) the bipartition corresponding to the left arrow above, (c) the bipartition corresponding to the right arrow above.

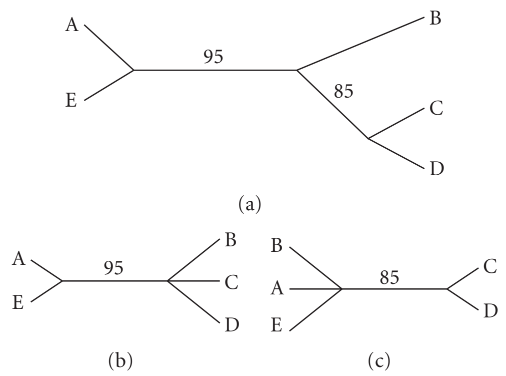


Figure 3: (a) Bootstrapped consensus tree with 5 taxa, (b) a bipartition with a 95% bootstrapped confidence value, (c) a bipartition with an 85% bootstrapped confidence value.

| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |

(a)

| | | | | |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 |

(b)

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 |

(c)

Figure 2: (a) A binary vector indexed by taxa names, (b) a binary representation of the bipartition in Figure 1(b), (c) a binary representation of the bipartition in Figure 1(c).

Let $t_1$, $t_2$, and $t_3$ be three different evolutionary trees of $n$ taxa and let $S(t_1)$, $S(t_2)$, and $S(t_3)$ be the respective spectra, then we say that $S(t_2)$ is more similar to $S(t_1)$ than $S(t_3)$ if $\|S(t_1) - S(t_2)\| < \|S(t_1) - S(t_3)\|$, here the operator $\| \cdot \|$ denotes the Euclidean distance between two points in bipartition space.

### 2.2. Representation of bipartitions

Let $A$ be a set of $n$ elements, and $b$ is a bipartition defined on a set $A$. Each bipartition $b$ splits a set $A$ into two subsets $m$ and its complement $m^C$, such that $A = m \cup m^C$.

We say that two bipartitions are *compatible* if there exists a tree whose spectrum includes both bipartitions. We say that two bipartitions are *conflicting* if they cannot appear in the same spectrum. In set notation, two bipartitions are compatible if a set (either $m$ or $m^C$) of one bipartition is a subset of one of the sets of the second bipartition; or, in other words, bipartitions $b_1$ and $b_2$ are compatible if and only if one of four possible conditions is satisfied:

$$(m_1 \subset m_2), (m_1 \subset m_2^C), (m_1^C \subset m_2), \text{ or } (m_1^C \subset m_2^C). \tag{1}$$

To handle bipartitions computationally in an efficient way, we can represent them effectively as binary masks. Figure 2(a) shows a binary vector indexed by the taxa in Figure 1(a). Figure 2(b) shows the binary representation of the bipartition in Figure 1(b) arbitrarily assigning 1 and 0 to the left and right bipartition, respectively. Figure 2(c) shows the binary representation of the bipartition in Figure 1(c).

Given our binary representation of bipartitions, there is a simple computation to test for compatibility between bipartitions. We say that two bipartitions are compatible if the following returns true:

$$\begin{aligned} ((b_1 \mid b_2) &== b_1)\ \|, \\ ((b_1 \mid b_2) &== b_2)\ \|, \\ ((b_1 \mid \sim b_2) &== b_1)\ \|, \\ ((b_1 \mid \sim b_2) &== \sim b_2)\ \|, \end{aligned} \tag{2}$$

where $b_1$ and $b_2$ denote bipartitions. Here the "|" operator represents the bitwise OR operation, the "∼" operator represents the bitwise negation, the "‖" operator represents the logical OR operation, and "==" represents the bitwise equality operator. Given the two masks from Figures 1(b) and 1(c), it is easy to see that they are compatible:

$$10001 \mid 11001 == 11001. \tag{3}$$

On the other hand, the bipartitions 11001 and 10011 are conflicting.

### 2.3. Consensus trees

It is customary to compute confidence values for the edges in an evolutionary tree via bootstrapping [7]. The computed tree represents a consensus tree over the bootstrap samples. The confidence values are typically chosen between 0 and 100. With this, a bipartition derived from a particular edge in the bootstrap consensus tree inherits the confidence value of that edge. This allows us to refine our spectrum vector notation, for example, $S(t_n) = (0, 67, 85, 0, \dots, 15, 0)$, where $t_n$ is now a bootstrapped consensus tree and the values in the vector represent the confidence values for the individual bipartitions.

Figure 3(a) shows a bootstrapped consensus tree with five taxa. The values on the edges represent the bootstrapped confidence values. Figures 3(b) and 3(c) show nontrivial bipartitions of the tree. Notice that the bipartitions inherit the confidence value of the edge that corresponds to the bipartition.

By computing a consensus tree on the bootstrap samples, it is possible to introduce biases due to the fact that phylogenies that do not agree with the plurality are suppressed. This is particularly critical in our case where the biases of this kind of computation might compound during an analysis. A different approach that avoids computing a consensus tree too early in an analysis is by taking advantage of the spectra of the bootstrap samples. Before we can describe this construction, we need to define what we mean by an *average spectrum*. Given $m$ spectra, $S_1, \dots, S_m$, in a bipartition space of $n$ taxa, we define the average spectrum $S_a$ as

$$S_a = \frac{1}{m}\sum_{k=1}^{m} S_k. \tag{4}$$

The summation of spectra is well defined as vector additions in bipartition space and the multiplication of a scalar and a vector simply scales the components of the vector.

The bootstrap approach can be summarized as follows.

(1) For the phylogenetic tree of each bootstrap sample, compute the corresponding spectrum.
(2) Compute the average spectrum $S_a$ over the bootstrap spectra.
(3) The values that appear in the vector for the average spectrum can now be interpreted as *confidence values*.

In step 3, we could multiply the average spectrum by 100 to make it compatible with the traditional bootstrap confidence values. A consequence of this approach is that the average spectrum is no longer guaranteed to represent a phylogenetic tree due to possible bipartition conflicts and this represents an extension of our definition of spectrum above that did not admit any conflicts. However, even in this extended definition of a spectrum we can retrieve a consensus tree from the average spectrum $S_a$ as follows:

(1) Sort the bipartitions in $S_a$ according to their confidence values.
(2) Delete all bipartitions in $S_a$ that conflict with more strongly supported bipartitions in $S_a$.
(3) Incrementally construct a consensus tree from the remaining bipartitions in $S_a$, starting with the bipartition with the strongest support to the bipartition with the weakest support.

Observe that computing the consensus tree for the average spectrum is a lossy operation (step 2) as before. However, the advantage of this approach is that we can defer this lossy operation as long as necessary. Note that we need only $n - 3$ top nonconflicting bipartitions. If conflicts are singular or minor events, they will not appear in the top $n - 3$ bipartitions because their confidence values will be low. If the conflicting bipartitions are among top $n - 3$, then the case deserves special attention. If the confidence values for bipartitions are rather small and randomly distributed over the data, this can serve as an indication that the data do not have a clean phylogenetic signal.

An interesting application of this is the construction of a consensus tree of multiple spectra in a bipartition space. If we interpret the spectra $S_1, \dots, S_m$ as a cluster in bipartition space, then the average spectrum can be viewed as the *centroid* of that cluster.

The following constructs a centroid consensus tree of $m$ given spectra, $S_1, \dots, S_m$.

(1) Compute $S_a$ for $S_1, \dots, S_m$
(2) Sort the bipartitions in $S_a$ according to their confidence values.
(3) Delete all bipartitions in $S_a$ that conflict with more strongly supported bipartitions in $S_a$.
(4) Incrementally construct a consensus tree from the remaining bipartitions in $S_a$, starting with the bipartition with the strongest support to the bipartition with the weakest support.

Note that this is essentially the same algorithm as above with the exception that the spectra, $S_1, \dots, S_m$ are not bootstrapped samples but arbitrary points in some bipartition space.

### 2.4. Unsupervised learning in bipartition space

Self-organizing maps [4] were introduced by Kohonen in 1982 and can be viewed as tools to visualize structure in high-dimensional data. Self-organizing maps are considered members of the class of unsupervised machine learning algorithms, since they do not require a predefined concept but will learn the structure of a target domain without supervision.

Typically, a self-organizing map consists of a rectangular grid of processing units. Multidimensional observations are represented as vectors. Each processing unit in the self-organizing map also consists of a vector called a reference vector or reference model. In our case, the multidimensional observations are spectra, where the number of possible bipartitions given $n$ taxa governs the dimensions of the spectra. The dimensions of processing elements of the map match the dimensionality of the observations.

The goal of the map is to assign values to the reference models on the map in such a way that all observations can be represented on the map with the smallest possible error. However, the map is constructed under constraints in the sense that the reference models cannot take on arbitrary values but are subject to a smoothing function called the neighborhood function. During training the values of the reference models on the map become ordered so that similar reference models are close to each other on the map and dissimilar ones are further apart from each other. This implies that similar observations will be mapped to similar regions on the map. Often reference models are referred to as centroids since they typically describe regions of observations with large similarities.

The training of the map is carried out by a sequential process, where $t = 1, 2, \dots$ is the step index. For each observation $\mathbf{x}(t)$ at time $t$, we first identify the index $c$ of some reference model which represents the best match in terms of Euclidean distance by the condition

$$c = \arg\min_i \|x(t) - m_i(t)\| \quad \forall i. \tag{5}$$

Here, the index $i$ ranges over all reference models on the map. The quantity $\mathbf{m}_i(t)$ refers to the reference model

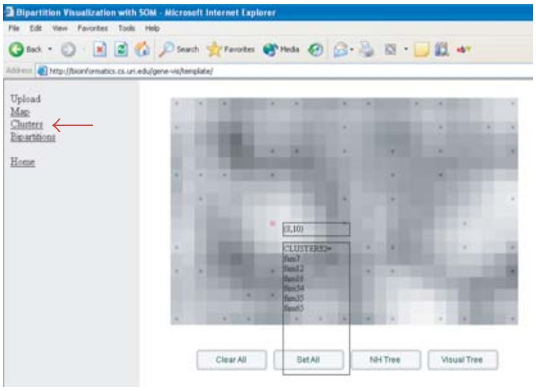


FIGURE 4: A typical visualization computed by GPX.

at position $i$ on the map at time step $t$. Next, all reference models on the map are updated with the following regression rule where model index $c$ is the reference model index as computed above:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}[\mathbf{x}(t) - \mathbf{m}_i(t)] \quad \forall i. \tag{6}$$

Here, $h_{ci}$ is the neighborhood function that is defined as follows:

$$h_{ci} = \begin{cases} 0 & \text{if } |c - i| > \beta, \\ \eta & \text{if } |c - i| \leq \beta, \end{cases} \tag{7}$$

where $|c - i|$ represents the distance between the best matching reference model at position $c$ and some other reference model at position $i$ on the map, $\beta$ is the neighborhood distance and $\eta$ is the learning rate. It is customary to express $\eta$ and $\beta$ also as functions of time. This computation is usually repeated over the available observations many times during the training phase of the map. Each iteration is called a training epoch.

An advantage of self-organizing maps is that they have an appealing visual representation. That is, the structure of the input domain is graphically represented as a 2-dimensional map. Figure 4 shows a typical map computed in GPX (here the map reconstructed from bipartition matrix of 14 Archaeal species).

Each square in the map represents a reference model. The shading of the map represents the level of quantization or mapping error for the map. Light shading represents a small quantization error; that is, the reference models in those areas match the observations very closely. Dark shading represents a large quantization error; that is, there is a poor match between reference models and observations. Contiguous areas of low quantization error represent clusters of similar entities. Figure 4 shows an interactive cluster layout of the GPX tool. Each cluster contains a set of orthologous families that we put together by the SOM algorithm. By moving a mouse pointer over the map, a user is able to highlight and select clusters of interest and reconstruct phylogenetic trees for the selection.

Here, we make use of this ability of self-organizing maps to visualize high-dimensional spaces in order to visualize similarities and dissimilarities of high-dimensional tree spectra. We would expect points in bipartition space that represent similar spectra to map close together on the visualization and vice versa. Once we have identified clusters of spectra, we can proceed to compute consensus trees for those clusters. Furthermore, we can now compare the trees calculated from individual clusters to the overall consensus tree, and we can investigate whether there exists substantial conflict between the bipartitions of various clusters. Furthermore, the clusters that result from this unsupervised learning allow the biologist to detect trends in the evolutionary histories of the participating genes which might provide insight into events such as horizontal transfers of individual genes or whole metabolic pathways. The fact that the spectra of individual gene families can be visualized as consensus trees and that it is possible to compute the average of several selected spectra and the corresponding majority consensus tree on the fly distinguishes our approach from other spectral approaches (e.g., [3, 8]).

### 2.5. The construction of gene families

One of the insights of recent evolutionary biology is that it is not sufficient to use one or a few genes to infer phylogenetic relationships among species. Therefore, we propose to use as many genes as possible in our analysis based on the notion of a *gene family*. A gene family is a collection of genes from different genomes that are related to each other and share a common ancestor. In general, a gene family may include both orthologs and paralogs [9]. Here, we consider only sets of putatively orthologous genes where each species contributes only one gene into a family. The evolutionary history of an individual gene family is a phylogenetic tree.

We select common gene families based on reciprocal best BLAST [10] hit criteria [11] with relaxation (see below). The reciprocal best BLAST hit method requires strong conservative relationships among the orthologs so that if a gene from species 1 selects a gene from species 2 as the best hit when performing a BLAST search with genome 1 against genome 2, then the gene 2 must in turn select gene 1 as the best hit when genome 2 is searched against genome 1. The requirement of reciprocity is very strict and often fails in the presence of paralogs. To select more orthologous sets, we relax the criteria of strict reciprocity by allowing a fixed number of broken connections.

The gene families are aligned with Clustalw version 1.83 using default parameters [12]. For each family, 100 bootstrapped replicates are generated and evaluated with the Phyml program [13] using the JTT model, four relative substitution rate categories, and an estimated shape parameter for the gamma distribution describing among site rate variation.

All 100 generated trees are split into their corresponding bipartition spectra and corresponding bootstrap support values are assigned to each bipartition by calculating how many times each bipartition is present in a family (the bootstrap procedure discussed in detail above). The result of these calculations is a spectrum for each gene family. Observe that trees calculated from individual bootstrap samples contain edges that are not part of a majority consensus tree, that is, the spectrum for a gene family can contain bipartitions that conflict with other bipartitions in the spectrum. For our purposes, this is important since it prevents information loss and avoids bias during our analyses.

We can now use the machinery developed above to investigate the consensus tree of the collection of gene families and whether there exist spectra that have a significant conflict with the overall consensus tree.

## 3. APPLICATION OF GPX

GPX, a tool based on the techniques developed above supports an active, investigation-style analysis where the user can interact with the visualization. The user is able to select centroids on the map and investigate consensus trees and conflicting bipartitions in the respective spectra. A detailed description of an experiment using GPX appears in [6]. In a first experiment, we analyzed 123 gene families of 14 archaea species. We found that sets of gene families exhibited substantial conflict with the overall organismal consensus tree corroborating findings of frequent gene transfers between organisms sharing the same or similar ecological niches [14, 15]. In the consensus over all 123 gene families, the representative of the Methanosarcinales (*Methanosarcina acetivorans*) grouped with the Haloarchaea (*Haloarcula marismortui and Halobacterium salinarum*) as expected from the analysis of ribosomal RNAs and enzymes involved in transcription and translation [16, 17]. Two clusters of gene families were recognized that strongly supported a conflicting bipartiton that places the homolog from *Methanosarcina acetivorans* with *Archaeoglobus fulgidus*. For one of these clusters, the relationships among the other archaea remained otherwise compatible with the consensus, suggesting gene transfer events between the ancestors of *Methanosarcina* and *Archaeoglobus*. However, in case of the second cluster formed by a single gene family, prolyl tRNA synthetases (prolylRS), the Haloarchaea grouped at the base of the euryarchaeota. This placement suggests that the ancestor of the Haloarchaea might have acquired this enzyme from outside the archaeal domain, a finding that was corraborated through more detailed phylogenetic analysis (Gogarten, unpublished). While the haloarchaeal prolylRS are more similar to bacterial than to archaeal homologs, database searches did not identify any sequence from an extant organism that is specifically related to the haloarchaeal prolyl tRNA synthetases. The donor of the haloarchaeal prolylRS is not a member of any of the bacterial or archaeal phyla that have prolylRS sequences in the current nonredundant or environmental databases; possibly the lineage that donated this enzyme has gone extinct as a distinct lineage, and only those genes that were donated to other lineages in the past survived into the presence [18]. These results were obtained by means of an originally developed interactive tool [6], which combines computationally expensive analysis of complex data with convenient visual representation of phylogenetic information.

## 4. CONCLUSIONS

We developed a comparative genomic analysis technique based on bipartition spectra and unsupervised learning. We have incorporated the techniques developed here into a web-based tool and have used this tool successfully in a set of analyses. The tool allows the user to reconstruct the evolutionary history shared by the plurality of gene family histories present within a collection of genomes; gene families with histories that are in conflict with the plurality are detected, and families which share conflicting histories can be recognized, thereby facilitating the discovery of major “highways of gene sharing” [15].

Bipartition spectrum analysis is not restricted to the SOM algorithm, other clustering algorithms, such as principal component analysis (PCA) [19] and local linear embedding (LLE) [20], can be applied to the analysis of large data sets. A new algorithm, generative topographic mapping (GTM) [21], displays maps similar to SOM but uses an expectation maximization (EM) algorithm instead of relying on neural network convergence. An alternative-to-traditional PCA is kernel PCA [22]. This algorithm is based on support vector machines, which allows it to easily deal with very wide datasets. ISOMAP [23] is an algorithm similar to LLE but distinguishes itself from LLE in that there is no need to solve a set of linear equations. To make comparative genomic studies a reality, we need to be able to include large numbers of genomes. This implies that we need to be able to handle large amounts of data. Future efforts will revolve around scaling up methodologies to include as many species as possible and testing different clustering algorithms for extraction of important phylogenetic information.

## ACKNOWLEDGMENTS

This work was in part supported through the NASA Applied Information System Research Program (NNG04GP90G). O. Zhaxybayeva was supported through a Postdoctoral Fellowship from the Canadian Institute of Health Research.

## REFERENCES

[1] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, “Prokaryotic evolution in light of gene transfer,” *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002.

[2] M. D. Hendy and D. Penny, “Spectral analysis of phylogenetic data,” *Journal of Classification*, vol. 10, pp. 5–24, 1993.

[3] O. Zhaxybayeva, P. Lapierre, and J. P. Gogarten, “Genome mosaicism and organismal lineages,” *Trends in Genetics*, vol. 20, no. 5, pp. 254–260, 2004.

[4] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 3rd edition, 2001.

[5] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, “Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples,” *DNA Research*, vol. 12, no. 5, pp. 281–290, 2005.

[6] N. Nahar, M. S. Poptsova, L. Hamel, and J. P. Gogarten, “GPX: a tool for the exploration and visualization of genome evolution,” in *Proceedings of the 7th IEEE International Symposium*

*on Bioinformatics & Bioengineering (BIBE '07)*, pp. 1338–1342, Boston, Mass, USA, October 2007.

[7] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[8] G. M. Lento, R. E. Hickson, G. K. Chambers, and D. Penny, "Use of spectral analysis to test hypotheses on the origin of pinnipeds," *Molecular Biology and Evolution*, vol. 12, no. 1, pp. 28–52, 1995.

[9] W. M. Fitch, "Homology: a personal view on some of the problems," *Trends in Genetics*, vol. 16, no. 5, pp. 227–231, 2000.

[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[11] O. Zhaxybayeva and J. P. Gogarten, "Boostrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses," *BMC Genomics*, vol. 3, p. 4, 2002.

[12] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[13] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[14] R. Jain, M. C. Rivera, J. E. Moore, and J. A. Lake, "Horizontal gene transfer accelerates genome innovation and evolution," *Molecular Biology and Evolution*, vol. 20, no. 10, pp. 1598–1602, 2003.

[15] R. G. Beiko, T. J. Harlow, and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14332–14337, 2005.

[16] C. Brochier, P. Forterre, and S. Gribaldo, "An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences," *BMC Evolutionary Biology*, vol. 5, p. 36, 2005.

[17] C. R. Woese, "Bacterial evolution," *Microbiology and Molecular Biology Reviews*, vol. 51, pp. 221–271, 1987.

[18] O. Zhaxybayeva and J. P. Gogarten, "Cladogenesis, coalescence and the evolution of the three domains of life," *Trends in Genetics*, vol. 20, no. 4, pp. 182–187, 2004.

[19] I. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2002.

[20] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, no. 2, pp. 119–155, 2004.

[21] C. M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: the generative topographic mapping," *Neural Computation*, vol. 10, no. 1, pp. 215–234, 1998.

[22] B. Scholkopf, A. Smola, and K. R. Muller, "Kernel principal component analysis," in *Advances in Kernel Methods-Support Vector Learning*, pp. 327–352, MIT press, Cambridge, Mass, USA, 1999.

[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 675741, 7 pages
doi:10.1155/2008/675741

*Research Article*

# A Scaffold Analysis Tool Using Mate-Pair Information in Genome Sequencing

**Pan-Gyu Kim,[1] Hwan-Gue Cho,[2] and Kiejung Park[1]**

[1] *SmallSoft Co., Ltd., Jang-Dong 59-5, Yusung-Gu, Daejeon 305-343, South Korea*
[2] *Department of Computer Science and Engineering, Pusan National University, Busan 609-735, South Korea*

Correspondence should be addressed to Kiejung Park, kjpark@smallsoft.co.kr

Received 1 September 2007; Accepted 24 December 2007

Recommended by Daniel Howard

We have developed a Windows-based program, *ConPath*, as a scaffold analyzer. *ConPath* constructs scaffolds by ordering and orienting separate sequence contigs by exploiting the mate-pair information between contig-pairs. Our algorithm builds directed graphs from link information and traverses them to find the longest acyclic graphs. Using end read pairs of fixed-sized mate-pair libraries, *ConPath* determines relative orientations of all contigs, estimates the gap size of each adjacent contig pair, and reports wrong assembly information by validating orientations and gap sizes. We have utilized ConPath in more than 10 microbial genome projects, including *Mannheimia succiniciproducens* and *Vibro vulnificus*, where we verified contig assembly and identified several erroneous contigs using the four types of error defined in *ConPath*. Also, *ConPath* supports some convenient features and viewers that permit investigation of each contig in detail; these include contig viewer, scaffold viewer, edge information list, mate-pair list, and the printing of complex scaffold structures.

## 1. INTRODUCTION

In 2001, the Human Genome Project (HGP) Consortium and Celera Genomics reported the first drafts of sequences of the human genome [1, 2]. The HGP Consortium used the hierarchical sequencing or "clone-by-clone" approach, whereas Celera Genomics used the whole genome shotgun (WGS) approach, which had been successfully used in 1995 to sequence the *H. influenzae* genome [3].

In the hierarchical sequencing approach, a tiling of large DNA sequences, such as bacterial artificial chromosome (BAC) or yeast artificial chromosome (YAC), are constructed for a genome, and each of the sequences is determined. The HGP Consortium used BAC as the large sequence, followed by shotgun sequencing of each BAC.

In sequencing the genome, owing to physical limitations of shotgun sequencing methods, the genome must be broken down into smaller portions, shotgun reads sized in the range of 600 bps (base-pairs) to 800 bps, and as the sequence data for each of these shotgun reads is produced, it must be connecting them with those adjacent and overlapping reads that have been previously sequenced, that is, to achieve an assembly of these smaller sequences into larger contiguous regions or "contigs."

In most cases, the sequences of shotgun reads are obtained by sequencing both ends of a DNA fragment whose approximate size is known. Such pair information, referred to as mate-pair information, constrains the placement of the reads within an assembly. In an ideal assembly, all read pairs are placed in such a manner as to satisfy the orientation and distance constraints imposed by the pairing. Mate-pair information can be used to determine the quality of an assembly, because most types of misassemblies lead to violations of these constraints.

In contrast to hierarchical sequencing, WGS breaks a whole genome into small pieces randomly, without shearing into large DNA pieces of intermediate size. WGS is faster and cheaper than hierarchical sequencing because of the simplicity of the processing steps. The success of WGS [4, 5] has increased its usage and the size of the genome to be sequenced has increased.

Although contig assembly programs are well established, less is known about scaffold analysis. While some of its features have been implemented to sequence specific genomes

[6–8], the features needed for general scaffold analysis and visualization have not been provided. *Consed* [9], a graphical tool for contig assembly, provides good visualization and helps to finish sequencing by connecting with *Autofinish* [10]; however, it does not have many features related to scaffold analysis.

It has been suggested that the contig scaffolding problem can be solved by *greedy-path merging algorithm* [8]. Moreover, *GigAssembler* can orient the contigs based on mRNA, paired plasmid ends, EST, and BAC end pairs [7].

This paper introduces a novel scaffold analysis tool, *ConPath*, which calculatesthe longest scaffolds. Due to the abundance of repeats in genomic DNA sequences, a purely overlap-based approach for WGS assembly is not feasible, but the use of mate-pair information is crucial. The *ConPath* program uses end read pairs of fixed-sized DNA libraries as mate-pairs to calculate orientations, orders, and gap sizes. It reads a *Phrap* [11] output file (∗.out) and an *ACE* format file, which contain contig structures and mate-pair information.

## 2. MATERIALS AND METHODS

### 2.1. Mate-pair information

The most important characteristic of *ConPath* is its ability to exploit the mate-pair information of large DNA fragments such as fosmids or cosmids, which are about 40 kbps(kilo base-pairs) in size, or BACs, which are about 100–300 kbps in size, rather than plasmids, which are about 2–10 kbps in size. Figure 1 shows an example of mate-pair end reads. A mate-pair is composed of two end reads that always face each other. Each end read, $b$ or $g$, has an orientation relative to the contig containing it. If the direction of an end read is the same as the direction of the contig, the former has direction $U$, otherwise, it has direction $C$. In Figure 1, $b$ has direction $U$ because the $C_l$ contig and $b$ read are in the same direction, whereas $g$ has direction $C$ because the $C_2$ contig and $g$ read are in opposite directions. The size of the mate-pair helps to estimate the gap size between contigs $C_1$ and $C_2$. When one contig contains one end of a mate-pair and a second contig contains the other end of the mate-pair, the two contigs are said to be linked by the mate-pair. A scaffold is a series of contigs that can be linked by mate-pairs. The connection relationship of all the contigs can be represented as a graph in which each contig is represented as a vertex. An edge is created between two contig vertices when they are linked by at least one mate-pair, and the number of linking mate-pairs between two contigs is defined as the edge weight.

### 2.2. Construction of scaffolds

To construct scaffolds using mate-pair information, a scaffold graph can be defined as follows.

Given a set of contigs $C = \{c_1, c_2, c_3, \ldots, c_n\}$, a mate-pair set $M = \{m_1, m_2, m_3, \ldots, m_l\}$, and a set of reads $R = \{r_1, r_2, r_3, \ldots, r_s\}$, let $G$ denote the scaffold graph using $C$ and $M$:

$$G = (C, E). \tag{1}$$

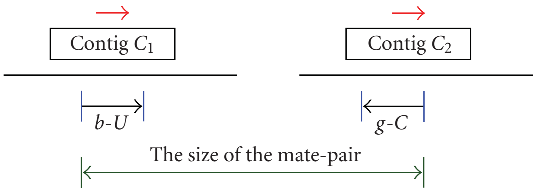


FIGURE 1: An example of mate-pair information. Mate-pair reads are indicated as read “$b$” and “$g$” and the relative directions to encompassing contigs are denoted as “$U$ (same direction)” and “$C$ (complementary direction).”

When a mate-pair $m_k = (r_i, r_j)$ exists, in which contig $c_s$ contains $r_i$ and contig $c_t$ contains $r_j$, there is an edge between contigs $c_s$ and $c_t$. Edge set $E$ is expressed as

$$E = \{e_{c_s c_t} \text{ iff } m_k = (r_i, r_j) \text{ exists for } r_i \in c_s, \\ r_j \in c_t, c_s \in C, \text{and } c_t \in C\}. \tag{2}$$

In constructing a scaffold graph, the linking level ($l$), the threshold value for the edge weights, was used as a filtering value in constructing and showing scaffolds on output. When an edge has a weight value smaller than the linking level ($l$), the edge is discarded from the graph.

Considering the errors that occur in base calling and contig assembly, the optimal construction of a scaffold graph is an NP-complete problem [8]. To practically solve this problem, *ConPath* uses a simple greedy algorithm. Whenever a new edge is added to the graph, graph $G$ is additive modified for that edge. This provides a feasible heuristic solution for a scaffold construction in linear time. Algorithm 1 shows the algorithm of *ConPath* to construct scaffolds.

### 2.3. Determination of the orders and orientations of contigs

It is worthwhile noting that *ConPath* determines the relative orientations of all contigs using the orientations of the end reads.

Figure 2 shows the determination of the order and orientations of three contigs using two mate-pairs. In Figure 2(a), $b_1$ and $g_1$ reads determine the relative orientation of contigs $C_1$ and $C_2$, and, in the same way, $b_2$ and $g_2$ reads determine the relative orientations of contigs $C_2$ and $C_3$ (see Figure 2(b)). The relative orientations of contigs $C_1, C_2$, and $C_3$ are determined by rotating the scaffold in Figure 2(b), as shown in Figure 2(c).

### 2.4. Estimation of the gap size between contigs

Assuming all mate-pairs have a fixed size, the size of the gap between two adjacent contigs is determined by the sizes of the two contigs and the positions of the end reads of contigs.

Suppose that contig $C_1$ contains $b$ read and contig $C_2$ contains g read. Let Gap $(C_1, C_2)$ be the gap size between $C_1$ and $C_2$. Let $P_s(b)$ and $P_e(b)$ be the start and end positions of $b$ read in $C_1$, respectively, and let $P_s(g)$ and $P_e(g)$ be the start

Table 1: Mate-pair information in real test datasets. The proportion of mate-pair reads for *V. vulnificus* is about double that for *M. succiniciproducens*.

| Genome | Genome length | Fold | Number of reads | Number of mate-pairs | Proportion of mate-pair reads relative to number of reads |
|---|---|---|---|---|---|
| *M. succiniciproducens* | 2.3 Mbp | 13.2 | about 25,000* | 275 | 2.2% |
| *V. vulnificus* | 5.1 Mbp | 11.7 | 76,971 | 1,781 | 4.5% |

* The numbers of reads for 4 versions of *M. succiniciproducens* show slight variation.

Table 2: Real test datasets. Four datasets for the *M. succiniciproducens* genome and one for the *V. vulnificus* genome were tested with *ConPath*. MP: mate-pair, MPIC: mate-pair in the same contigs.

| Data name | Number of contigs | Number of MPs | Number of MPICs | Average size of MP(fosmid)s |
|---|---|---|---|---|
| MH1 | 98 | 238 | 72 | 37,673 bp |
| MH2 | 86 | 240 | 115 | 38,102 bp |
| MH3 | 85 | 240 | 120 | 38,157 bp |
| MH4 | 112 | 240 | 108 | 37,917 bp |
| VV | 334 | 1,220 | 454 | 33,024 bp |

```
MakeScaffold(mate-pair set)
{
   initial scaffold graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}; \mathbf{V} = \{\text{all contigs}\}, \mathbf{E} = \{\})$
   assign each mate-pair to corresponding an edge → edge set $\mathbf{E}$
   remove self-collision mate-pairs
   while (edge set $\mathbf{E}$ is not empty)
        find an edge with maximum weight from $\mathbf{E} \rightarrow \mathbf{e}(\mathbf{k}, 1)$
        if ($\mathbf{e}(\mathbf{k}, 1)$ does not conflict with $\mathbf{G}$)
            add $\mathbf{e}(\mathbf{k}, 1)$ to graph $\mathbf{G}$
        delete $\mathbf{e}(\mathbf{k}, 1)$ from edge set $\mathbf{E}$
   end while
}
```

Algorithm 1: The algorithm for scaffold construction. *ConPath* uses a simple greedy algorithm to obtain a feasible heuristic solution for an NP-complete problem.

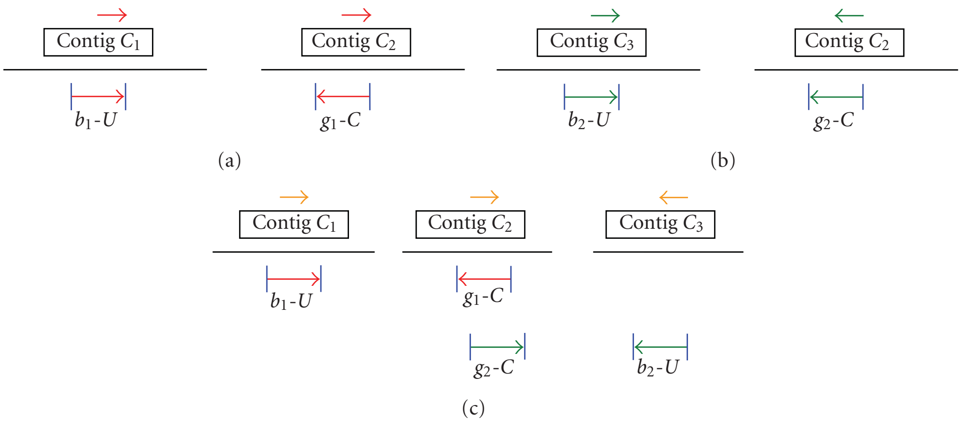


Figure 2: Determining the relative orientations of contigs using mate-pair information. (a): $b_1$ and $g_1$ reads determine the relative orientation of contigs $C_1$ and $C_2$; (b): $b_2$ and $g_2$ reads determine the relative orientations of contigs $C_2$ and $C_3$; and (c): the relative orientations of contigs $C_1$, $C_2$, and $C_3$ are determined by rotating the scaffold in Figure 2(b).

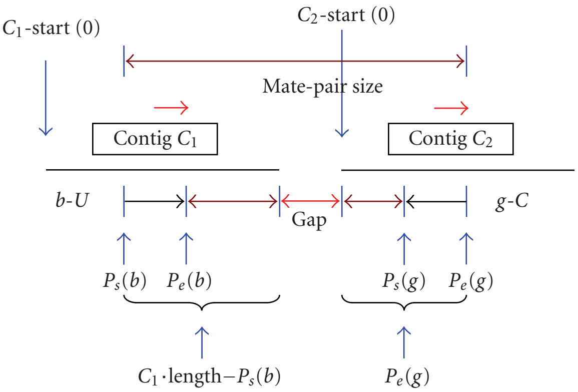
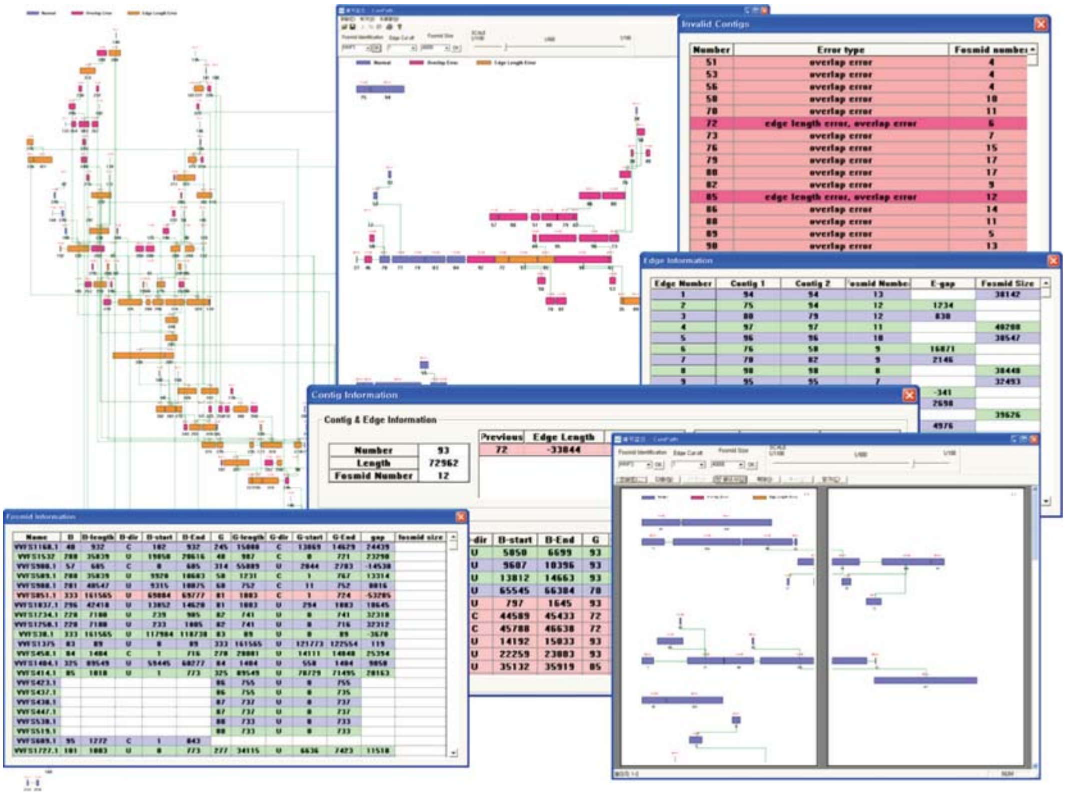


FIGURE 3: Estimation of the gap size between contigs when $b$ has direction $U$ and $g$ has direction $C$. The gap size between $C_1$ and $C_2$ can be calculated as mate_pair size $- \{(C_1 \cdot \text{length} - P_s(b)) + P_e(g)\}$.



Invalid Contigs

| Number | Error type | Fosmid number |
|---|---|---|
| 51 | overlap error | 4 |
| 53 | overlap error | 4 |
| 56 | overlap error | 4 |
| 58 | overlap error | 10 |
| 70 | overlap error | 11 |
| 72 | edge length error, overlap error | 6 |
| 73 | overlap error | 7 |
| 76 | overlap error | 15 |
| 79 | overlap error | 17 |
| 80 | overlap error | 17 |
| 82 | overlap error | 9 |
| 85 | edge length error, overlap error | 12 |
| 86 | overlap error | 14 |
| 88 | overlap error | 11 |
| 89 | overlap error | 5 |
| 90 | overlap error | 13 |

Edge Information

| Edge Number | Contig 1 | Contig 2 | Fosmid Number | E-gap | Fosmid Size |
|---|---|---|---|---|---|
| 1 | 94 | 94 | 13 | | 38142 |
| 2 | 75 | 94 | 12 | 1234 | |
| 3 | 80 | 79 | 12 | 830 | |
| 4 | 97 | 97 | 11 | | 40700 |
| 5 | 96 | 96 | 10 | | 38547 |
| 6 | 76 | 58 | 9 | 16871 | |
| 7 | 70 | 82 | 9 | 2146 | |
| 8 | 98 | 98 | 8 | | 38448 |
| 9 | 95 | 95 | 7 | | 32493 |
| | | | | -341 | |
| | | | | 2698 | |
| | | | | | 39626 |
| | | | | 4976 | |

Contig Information

Contig & Edge Information

| Number | 93 |
|---|---|
| Length | 72962 |
| Fosmid Number | 12 |

| Previous | Edge Length |
|---|---|
| 72 | -33844 |

| dir | B-start | B-End | G |
|---|---|---|---|
| U | 5050 | 6699 | 93 |
| U | 9607 | 10396 | 93 |
| U | 13812 | 14663 | 93 |
| U | 65545 | 66384 | 70 |
| U | 797 | 1645 | 93 |
| C | 44589 | 45433 | 72 |
| C | 45788 | 46638 | 72 |
| U | 14192 | 15033 | 93 |
| U | 22259 | 23083 | 93 |
| U | 35132 | 35919 | 85 |

Fosmid Information

| Name | B | B-length | B-dir | B-start | B-End | G | G-length | G-dir | G-start | G-End | gap | fosmid size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VVFS1168.1 | 40 | 932 | C | 182 | 932 | 245 | 15000 | C | 13869 | 14629 | 24439 | |
| VVFS1532 | 280 | 35839 | U | 19050 | 20616 | 40 | 987 | C | 0 | 721 | 23290 | |
| VVFS900.1 | 57 | 685 | C | 0 | 685 | 314 | 55009 | U | 2044 | 2783 | -14530 | |
| VVFS509.1 | 280 | 35839 | U | 9920 | 10683 | 58 | 1231 | C | 1 | 767 | 13314 | |
| VVFS908.1 | 281 | 40547 | U | 9315 | 10075 | 68 | 752 | C | 11 | 752 | 8016 | |
| VVFS851.1 | 333 | 161565 | U | 69004 | 69777 | 81 | 1083 | C | 1 | 724 | -53285 | |
| VVFS1837.1 | 296 | 42418 | U | 13852 | 14620 | 81 | 1083 | U | 294 | 1083 | 10645 | |
| VVFS1734.1 | 220 | 7100 | U | 239 | 985 | 82 | 741 | U | 0 | 741 | 32318 | |
| VVFS1258.1 | 220 | 7100 | U | 233 | 1005 | 82 | 741 | U | 0 | 716 | 32312 | |
| VVFS38.1 | 333 | 161565 | U | 117984 | 118738 | 83 | 89 | U | 0 | 89 | -3670 | |
| VVFS1375 | 83 | 89 | U | 0 | 89 | 333 | 161565 | U | 121773 | 122554 | 119 | |
| VVFS458.1 | 84 | 1484 | C | 1 | 716 | 278 | 28081 | U | 14111 | 14848 | 25394 | |
| VVFS1484.1 | 325 | 89549 | U | 59445 | 60277 | 84 | 1484 | U | 558 | 1484 | 9850 | |
| VVFS414.1 | 85 | 1818 | U | 1 | 773 | 325 | 89549 | U | 70729 | 71495 | 28163 | |
| VVFS423.1 | | | | | | 86 | 755 | U | 0 | 755 | | |
| VVFS437.1 | | | | | | 86 | 755 | U | 0 | 735 | | |
| VVFS438.1 | | | | | | 87 | 737 | U | 0 | 737 | | |
| VVFS447.1 | | | | | | 87 | 737 | U | 0 | 737 | | |
| VVFS538.1 | | | | | | 88 | 733 | U | 0 | 733 | | |
| VVFS519.1 | | | | | | 88 | 733 | U | 0 | 733 | | |
| VVFS689.1 | 95 | 1272 | C | 1 | 843 | | | | | | | |
| VVFS1727.1 | 181 | 1083 | U | 0 | 773 | 277 | 34115 | U | 6636 | 7423 | 11518 | |

FIGURE 4: A set of snapshots of *ConPath*. *ConPath* provides a set of useful information, "mate-pair information", "edge information", "contig path", and "invalid contigs" by checking for the 4 types of error.

and end positions of $g$ read in $C_2$, respectively. Considering all the possible directions of a mate-pair of two end reads, *ConPath* estimates the gap size as

*b-U and g-U*: Gap $(C_1, C_2)$ = mate_pair size $- \{(C_1 \cdot \text{length} - P_s(b)) + (C_2 \cdot \text{length} - P_s(g))\}$

*b-U and g-C*: Gap $(C_1, C_2)$ = mate_pair size $- \{(C_1 \cdot \text{length} - P_s(b)) + P_e(g))\}$

*b-C and g-U*: Gap $(C_1, C_2)$ = mate_pair size $- \{P_e(b) + (C_2 \cdot \text{length} - P_s(g))\}$

*b-C and g-C*: Gap $(C_1, C_2)$ = mate_pair size $- \{P_e(b) + P_e(g)\}$

Figure 3 shows the procedure for estimating the gap size between contigs when $b$ and $g$ have $U$ and $C$ directions,

TABLE 3: Number of reported errors in scaffold construction for 5 dataset.

| Data name | | $l$ | | | |
|---|---|---|---|---|---|
| | Errors* | 1 | 2 | 3 | 4 |
| MH1 | Self Collision | 0 | 0 | 0 | 0 |
| | Gap size | 3 | 3 | 3 | 0 |
| | Overlap | 22 | 2 | 0 | 2 |
| MH2 | Self collision | 2 | 2 | 2 | 2 |
| | Gap size | 2 | 2 | 2 | 2 |
| | Overlap | 20 | 2 | 2 | 0 |
| MH3 | Self collision | 0 | 0 | 0 | 0 |
| | Gap size | 5 | 0 | 0 | 0 |
| | Overlap | 18 | 0 | 0 | 0 |
| MH4 | Self collision | 0 | 0 | 0 | 0 |
| | Gap size | 0 | 0 | 0 | 0 |
| | Overlap | 0 | 0 | 0 | 0 |
| VV | Self collision | 16 | 16 | 10 | 7 |
| | Gap size | 65 | 7 | 3 | 2 |
| | Overlap | 85 | 24 | 0 | 4 |

* Mate-pair size errors were excluded because these errors do not depend on $l$.

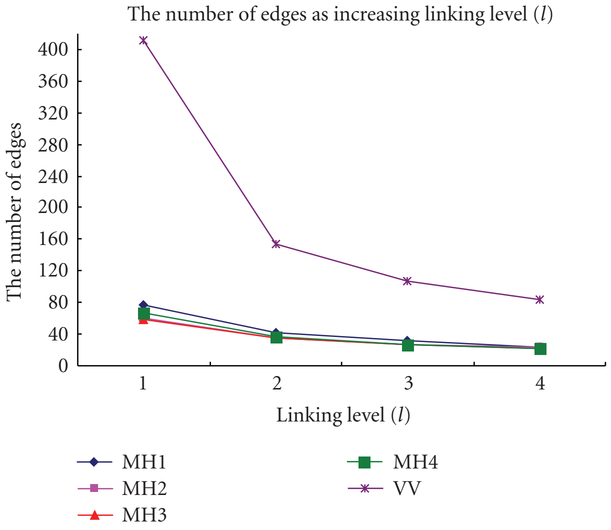


FIGURE 5: Distribution of the number of edges according to linking level ($l$). *ConPath* constructed the best scaffolds at linking level 2 while minimizing edge loss.

respectively. The orientations of contigs $C_1$ and $C_2$ are set in the same direction. The length of part of the mate-pair library in contig $C_1(C_1 \cdot \text{length} - P_s(b))$ and the length of part of the mate-pair library in contig $C_2(P_e(g))$ are calculated. Finally, the gap size is calculated as

$$\text{mate\_pair size} - \{(C_1 \cdot \text{length} - P_s(b)) + P_e(g)\}. \qquad (3)$$

### 2.5. Detection of erroneous contigs

One important feature of *ConPath* is the verification of a contig assembly by identifying erroneous contigs. We have defined 4 types of contig assembly errors to check the quality of a contig assembly.

#### Self-collision error

When the number of mate-pairs connecting two adjacent contigs is more than 2, and there is an inconsistency in determining the orientation of contigs with mate-pairs, the error is defined as a self-collision error, the most serious error type. If this error occurs, the contigs should be inspected manually one by one.

#### Mate-pair size error

When a mate-pair of an end read is contained in a contig, the real size of this mate-pair can be calculated. If the difference between the calculated and predefined sizes is larger than a threshold value, the error is defined as a mate-pair size error. This type of error is very critical to the contig assembly process.

#### Gap-size error

If the gap size between two contigs is a negative value, it indicates that the two contigs should be merged in the contig assembly process; this is defined as a gap size error.

#### Overlap error

After calculating the distances of all adjacent contigs, any two nonadjacent contigs can be overlapped due to the accumulation of errors in gap size estimations. This type of error is defined as an overlap error, which happens rarely and is not so critical.

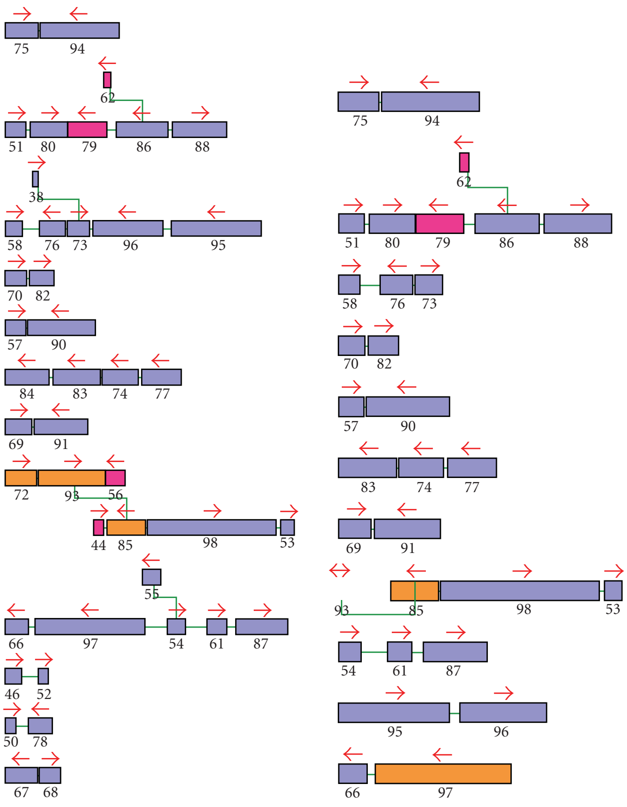
Identifying error types is useful in verifying and correcting the final result of a contig assembly. If a contig has more than two types of errors, it is highly probable that a misassembled contig is present. *ConPath* assigns different colors to contigs by the number of error types, with nonerroneous contigs colored blue. When one contig has more than one error, *ConPath* assigns this contig a reddish color, with the intensity proportional to the number of error types. Therefore, we can check the quality of the final result of a contig assembly by simply inspecting the color information in the scaffold visualization window of *ConPath*.

### 2.6. Implementation

*ConPath* was implemented on a Windows XP system using Visual C++. It provides a user-friendly interface and shows visual and color-informative outputs, which can help analyze scaffolds both intuitively and informatively. *ConPath* provides dialogue windows for "mate-pair information", "edge information", "contig path", and "invalid contigs" by automatically checking for the 4 types of errors. Scaffolds are displayed graphically in proportion to the real sizes of vertices and edges after aligning vertices and edges to avoid graphical collision, and the detailed information for each vertex and edge is shown on a pop-up window. *ConPath*

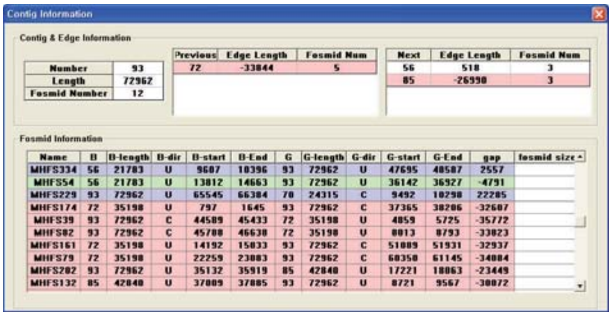
Table 4: Comparison of *ConPath* with other scaffold tools.

| Comparison item | Tools | | | |
|---|---|---|---|---|
| | *ConPath* | *Consed* | *Autofinish* | *Bambus* |
| Accuracy of scaffold | Medium | Medium | Medium | Strong |
| Construction time | Strong | Strong | Strong | Strong |
| Visualization | Strong | Medium | Weak | Weak |
| Error detection | Strong | Medium | Medium | Medium |
| Additional information | Strong | Strong | Medium | Medium |



(a)

(b)



| | |
|---|---|
| Number | 93 |
| Length | 72962 |
| Fosmid Number | 12 |

| Previous | Edge Length | Fosmid Num |
|---|---|---|
| 72 | -33844 | 5 |

| Next | Edge Length | Fosmid Num |
|---|---|---|
| 56 | 518 | 3 |
| 85 | -26990 | 3 |

| Name | B | B-length | B-dir | B-start | B-End | G | G-length | G-dir | G-start | G-End | gap | fosmid size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MHFS334 | 56 | 21783 | U | 9607 | 10396 | 93 | 72962 | U | 47695 | 48587 | 2557 | |
| MHFS54 | 56 | 21783 | U | 13812 | 14663 | 93 | 72962 | U | 36142 | 36927 | -4791 | |
| MHFS229 | 93 | 72962 | U | 65545 | 66384 | 78 | 24315 | C | 9492 | 10298 | 22285 | |
| MHFS174 | 72 | 35198 | U | 797 | 1645 | 93 | 72962 | C | 37365 | 38206 | -32687 | |
| MHFS39 | 93 | 72962 | C | 44589 | 45433 | 72 | 35198 | U | 4859 | 5725 | -35772 | |
| MHFS82 | 93 | 72962 | C | 45788 | 46638 | 72 | 35198 | U | 8013 | 8793 | -33823 | |
| MHFS161 | 72 | 35198 | U | 14192 | 15033 | 93 | 72962 | C | 51089 | 51931 | -32937 | |
| MHFS79 | 72 | 35198 | U | 22259 | 23083 | 93 | 72962 | C | 60350 | 61145 | -34084 | |
| MHFS202 | 93 | 72962 | U | 35132 | 35919 | 85 | 42840 | U | 17221 | 18063 | -23449 | |
| MHFS132 | 85 | 42840 | U | 37089 | 37885 | 93 | 72962 | U | 8721 | 9567 | -30072 | |

(c)

Figure 6: An example of the detection of mis-assembled contigs. (a): Scaffolds for MH1 at linking level 2; (b): scaffolds for MH1 at linking level 3; (c): information on contig 93.

can produce a large picture for all scaffolds by assembling separately printed module pictures. Figure 4 shows various viewers and dialogues of *ConPath*.

## 3. EXPERIMENTS AND DISCISSION

We tested *ConPath* using both artificial and real data. Artificial data were generated in two different versions: *R* (randomly) and *U* (uniformly). The *R* version consisted of contigs of random sizes, whereas the *U* version consisted of contigs of uniform size. In these artificial data experiments, *ConPath* showed very successful scaffold constructions using mate-pair information. From experiments with artificial data, *ConPath* made a reasonable scaffold construction in linear time.

*ConPath* worked very successfully and efficiently on real data sets, in sequencing the *Mannheimia succiniciproducens* and *Vibro vulnificus* genomes. *ConPath* verified the results of contig assembly by detecting misassembled contigs. Table 1vspace1pt shows the mate-pair information in these real datasets. Four datasets were tested in sequencing the *M. succiniciproducens* genome, whereas one dataset was tested in sequencing the *V. vulnificus* genome, to verify the results of contig assem-bly. Table 2 shows these results. MH1, MH2, MH3, and MH4 are the contig assembly results of the *M. succiniciproducens* genome and VV is the contig assembly result for the *V. vul-nificus* genome. For the *M. succiniciproducens* genome, going from MH1 to MH4 increased the reliability of the contig assembly results.

We examined the edge number according to linking level (see Figure 5). *ConPath* was most successful at linking level 2 by minimizing the loss of edges.

Table 3 shows the detected errors in scaffold construction for the 5 datasets. Among the *M. succiniciproducens* datasets, MH1 had the most errors, whereas MH4 had no erroneous contigs. These results show that identifying the 4 types of errors for contigs is effective in verifying the result of contig assembly.

Figure 6 shows the constructed scaffolds at linking levels 2 and 3 for the MH1 dataset. Contig 93 is suspected of being erroneous because it has several erroneous contigs on both sides. *ConPath* showed that contig 93 was misassembled. The contig information dialogue box for contig 93 is shown in Figure 6(c).

Table 4 shows a comparison of features of several scaffold analysis tools, including *ConPath*, *Consed* [9], *Autofinish* [10], and *Bambus* [12]. Compared with these other tools, *ConPath* has very good features for 5 criteria. Most importantly, *ConPath* helps users to intuitively verify the contig assembly by providing many visualization features and additional information to detect erroneous contigs.

## 4. CONCLUSION

A scaffold analyzer is a very important tool in genome sequencing, in that it can verify the results of contig assembly and to identify misassembled contigs. We have developed *ConPath*, a scaffold analyzer that exploits mate-pair information to construct scaffolds by ordering and orienting separate sequence contigs. *ConPath* provides various useful viewers and dialogue boxes for intuitive understanding. Using end read pairs of a fixed-sized mate-pair library, *ConPath* candetermine the relative orientations of all contigs successfully, and estimate the gap size of each adjacent contig pair. We defined 4 types of errors to detect misassembly. *ConPath* was used successfully in sequencing several microbial genomes, including the *M. succiniciproducens* genome [13]. *ConPath* is, therefore, a useful scaffold analyzer to verify contig assembly by detecting erroneous contigs.

*ConPath* will doubtless improve as its algorithm becomes more correct and efficient, as well as through the development of additional features, such as primer design for the finishing step and a sequence read viewer.

## REFERENCES

[1] E. S. Lander, L. M. Linton, B. Birren, et al., “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, pp. 860–921, 2001.

[2] J. C. Venter, M D. Adams, E. W. Myers, et al., “The sequence of the human genome,” *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[3] R. D. Fleischmann, M. D. Adams, O. White, et al., “Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*,” *Science*, vol. 269, no. 5223, pp. 496–512, 1995.

[4] E. W. Myers, “Whole-genome DNA sequencing,” *Computing in Science ’ Engineering*, vol. 1, no. 3, pp. 33–43, 1999.

[5] J. L. Weber and E. W. Myers, “Human whole-genome shotgun sequencing,” *Genome Research*, vol. 7, no. 5, pp. 401–409, 1997.

[6] V. Magrini, W. C. Warren, J. Wallis, et al., “Fosmid-based physical mapping of the *Histoplasma capsulatum* genome,” *Genome Research*, vol. 14, no. 8, pp. 1603–1609, 2004.

[7] W. J. Kent and D. Haussler, “Assembly of the working draft of the human genome with GigAssembler,” *Genome Research*, vol. 11, no. 9, pp. 1541–1548, 2001.

[8] D. H. Huson, K. Reinert, and E. W. Myers, “The greedy path-merging algorithm for contig scaffolding,” *Journal of the ACM*, vol. 49, no. 5, pp. 603–615, 2002.

[9] D. Gordon, C. Abajian, and P. Green, “Consed: a graphical tool for sequence finishing,” *Genome Research*, vol. 8, no. 3, pp. 195–202, 1998.

[10] D. Gordon, C. Desmarais, and P. Green, “Automated finishing with autofinish,” *Genome Research*, vol. 11, no. 4, pp. 614–625, 2001.

[11] “*Phrap*,” . http://www.genome.washington.edu/UWGC/ analysis tools/Phrap/htm.

[12] M. Pop, D. S. Kosack, and S. L. Salzberg, “Hierarchical scaffolding with Bambus,” *Genome Research*, vol. 14, no. 1, pp. 149–159, 2004.

[13] S. H. Hong, J. S. Kim, S. Y. Lee, et al., “The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*,” *Nature Biotechnology*, vol. 22, no. 10, pp. 1275–1281, 2004.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 518093, 10 pages
doi:10.1155/2008/518093

*Research Article*

# Fast Parallel Molecular Algorithms for DNA-Based Computation: Solving the Elliptic Curve Discrete Logarithm Problem over $GF(2^n)$

**Kenli Li,[1,2] Shuting Zou,[1] and Jin Xv[2]**

[1] *Embedded System and Networking Laboratory, College of Computer and Communication, Hunan University, Changsha 410082, China*
[2] *Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*

Correspondence should be addressed to Shuting Zou, zst991221@163.com

Received 16 July 2007; Accepted 25 January 2008

Recommended by Daniel Howard

Elliptic curve cryptographic algorithms convert input data to unrecognizable encryption and the unrecognizable data back again into its original decrypted form. The security of this form of encryption hinges on the enormous difficulty that is required to solve the elliptic curve discrete logarithm problem (ECDLP), especially over $GF(2^n)$, $n \in Z^+$. This paper describes an effective method to find solutions to the ECDLP by means of a molecular computer. We propose that this research accomplishment would represent a breakthrough for applied biological computation and this paper demonstrates that in principle this is possible. Three DNA-based algorithms: a parallel adder, a parallel multiplier, and a parallel inverse over $GF(2^n)$ are described. The biological operation time of all of these algorithms is polynomial with respect to $n$. Considering this analysis, cryptography using a public key might be less secure. In this respect, a principal contribution of this paper is to provide enhanced evidence of the potential of molecular computing to tackle such ambitious computations.

## 1. INTRODUCTION

This paper proposes theoretical work that introduces powerful algorithms of molecular computation that could potentially compromise the security that is afforded by certain cryptography algorithms. Molecular computation [1] involves biochemistry and DNA rather than silicon chips to tackle formidable computations. Theoretical aspects of this interdisciplinary field are important to develop the potential and interest in this form of computation [2].

Elliptic curve cryptography (ECC) is a mathematical approach to public key cryptography using elliptic curves that are typically defined over finite fields [3]. Elliptic curves [4, 5] constitute a major area of current research that is particularly important to number theory, for example, elliptic curves had a role in the recent proof of Fermats last theorem. As applied to cryptography, not only has ECC become applied in Diffie-Hellman key exchange but also in the digital signature algorithm (DSA), a US federal government standard for digital signatures. It is known as the elliptic curve DSA (ECDSA) or that variant of the DSA operating on elliptic curve groups.

The security of these cryptosystems relies on the difficulty of solving the elliptic curve discrete logarithm problem [6, 7]. If $P$ is a point with order $m$ on an elliptic curve, and $Q$ is some other point on the same curve, then the elliptic curve discrete logarithm problem is to determine an $l$ such that $Q = lP$, where $l$ is an integer and $0 \leq l \leq m - 1$. If this problem can be solved efficiently, then elliptic curve-based cryptosystems can be broken efficiently.

In order to tackle such a problem, Feynman proposed molecular computation in 1961 [8]. However, his idea was not implemented experimentally for some decades. In 1994, Adleman succeeded in solving an instance of the Hamiltonian path problem in a test tube, simply by the manipulation of DNA strands [1]. Following this, Lipton demonstrated that the Adleman techniques offered a solution to the satisfiability problem (the first considered NP-complete problem) [9].

Recent advances in molecular biology [10, 11] have made it possible to produce roughly $10^{18}$ DNA strands in a test tube. Those $10^{18}$ DNA strands can be made to represent $10^{18}$ bits of information. In a distant future, if biological operations may be run error free using a test tube with $10^{18}$ DNA strands, then it would be possible to process $10^{18}$ bits of information simultaneously. More details about test tube distributed systems are given in [2]. The objective for biological computing technology is to provide this enormous amount of parallelism for dealing with computationally intensive real world problems [12–14].

Advancement in DNA computing has already been made in many areas. In the field of cryptology, Boneh et al. have cracked DES using identical principles to those of Adleman's solution of the travelling salesman problem. Also, Chang et al. have developed a way to factor integers. They proposed three DNA-based algorithms: parallel subtractor; parallel comparator; and parallel modular arithmetic unit [15, 16].

In this paper, we take a step further with respect to Chang's work [16] in order to solve the elliptic curve discrete logarithm problem. We develop DNA-based algorithms for a parallel adder; a parallel multiplier; a parallel divider over $GF(2^n)$ (i.e., a Galois field of characteristic 2); and a parallel adder for adding points on elliptic curves. We accomplish all of these by means of basic biological operations. We also showed that cryptosystems based on elliptic curves can be broken. Our work presents clear evidence of molecular computing abilities to accomplish complex mathematical operations.

The paper is organized as follows. Section 2 gives a brief background on DNA computing. Section 3 introduces the DNA computing that solves the elliptic curve discrete logarithm problem, for solution spaces of DNA strands. Conclusions are drawn in the final section.

## 2. BACKGROUND

DNA (*DeoxyriboNucleic Acid*) is the *molecule* that plays the main role in DNA-based computing. DNA is a polymer, which is strung together from monomers called *deoxyriboNucleotides*. Distinct nucleotides are detected only with their bases, which come in two sorts: *purines* and *pyrimidines*. Purines include *adenine* and *guanine*, abbreviated $A$ and $G$. Pyrimidines contain *cytosine* and *thymine*, abbreviated $C$ and $T$. A DNA strand is essentially a sequence (polymer) of four types of nucleotides detected by one of four bases they contain. Two strands of DNA can form (under appropriate conditions) a double strand, if the respective bases are the Watson-Crick complements of each other—$A$ matches $T$ and $C$ matches $G$. Hybridization is a special technology term for the pairing of two single DNA strands to make a double helix and also takes advantages of the specificity of DNA base pairing for the detection of specific DNA strands (for more discussions of the relevant biological background, refer to [10, 11]).

In the past decade, there have been revolutionary advances in the field of biomedical engineering, particularly in recombinant DNA and RNA manipulating. Due to the industrialization of the biotechnology field, laboratory techniques for recombinant DNA and RNA manipulation are becoming highly standardized. Basic principles about recombinant DNA can be found in [17–20]. In the following, we describe five biological operations that are useful for solving the elliptic curve discrete logarithm problem.

A (test) tube is a set of molecules of DNA (a multiset of finite strings over the alphabet $\{A, C, G, T\}$). Given a tube, one can perform the following operations.

(1) *Extract*. Given a tube $P$ and a short single strand of DNA, $S$, the operation produces two tubes $+(P, S)$ and $-(P, S)$, where $+(P, S)$ is all of the molecules of DNA in $P$ which contain $S$ as a substrand and $-(P, S)$ is all of the molecules of DNA in $P$ which do not contain $S$.

(2) *Merge*. Given tubes $P_1$ and $P_2$, yield $\cup(P_1, P_2)$, where $\cup(P_1, P_2) = P_1 \cup P_2$. This operation is to pour two tubes into one, without any change in the individual strands.

(3) *Amplify*. Given a tube $P$, the operation Amplify $(P, P_1, P_2)$, will produce two new tubes $P_1$ and $P_2$ so that $P_1$ and $P_2$ are totally a copy of $P$ ($P_1$ and $P_2$ are now identical) and $P$ becomes an empty tube.

(4) *Append*. Given a tube $P$ containing a short strand of DNA $Z$, the operation will append $Z$ onto the end of every strand in $P$.

(5) *Append-head*. Given a tube $P$ containing a short strand of DNA, $Z$, the operation will append $Z$ onto the head of every strand in $P$.

## 3. FINDING THE DISCRETE LOGARITHM ON ELLIPTIC CURVE OVER $GF(2^n)$

### 3.1. *Elliptic curve public key cryptosystem over* $GF(2^n)$

An elliptic curve is defined to be the set of solutions $(x, y) \in GF(2^n) \times GF(2^n)$ to the equation

$$y^2 + xy = x^3 + ax^2 + b, \tag{1}$$

where $a, b \in GF(2^n)$ and $b \neq 0$, together with the point on the curve at infinity $O$, (with homogeneous coordinates $(0, 0)$).

The points on an elliptic curve form an Abelian group under a well-defined group operation. The identity of the group operation is the point $O$. For $P = (x_1, y_1)$ a point on the curve, we define $-P$ to be $(x_1, y_1 + x_1)$, so $P + (-P) = (-P) + P = O$. Now suppose $P$ and Q are not $O$, and $P \neq -Q$. Let $P$ be as above and $Q = (x_2, y_2)$, then $P + Q = (x_3, y_3)$, where

$$\begin{gathered} x^3 = \mu^2 + \mu + x_1 + x_2 + a, \\ y_3 = \mu(x_1 + x_3) + x_3 + y_{1,} \\ \mu = \begin{cases} \dfrac{y_2 + y_1}{x_2 + x_1} & \text{if } P \neq Q, \\ \dfrac{x_1^2 + y}{x_1} & \text{if } P = Q, \end{cases} \end{gathered} \tag{2}$$

(refer to [21]).

In this paper, for convenience, use $b_{n-1}b_{n-2} \cdots b_1b_0$ to denote the value of $b_{n-1}\omega^{n-1} + b_{n-2}\omega^{n-2} + \cdots + b_1\omega + b_0$ over $GF(2^n)$.

```
(1) For $j = 0$ to $n - 1$
    (1a) $T_1 = +(T_0, x^1_{p+j})$ and $T_2 = -(T_0, x^1_{p+j})$
    (1b) $T_3 = +(T_1, x^1_{q+j})$ and $T_4 = -(T_1, x^1_{q+j})$
    (1c) $T_5 = +(T_2, x^1_{q+j})$ and $T_6 = -(T_2, x^1_{q+j})$
    (1d) $T_7 = \cup(T_4, T_5)$ and $T_8 = \cup(T_3, T_6)$
    (1e) Append $(T_7, x^1_{r+j})$ and Append $(T_8, x^0_{r+j})$
    (1f) $T_0 = \cup(T_7, T_8)$
    EndFor
EndProcedure
```

Algorithm 1: Procedure ParallelAdder $(T_0, n, p, q, r)$.

Let $E$ be an elliptic curve defined over $GF(2^n)$, and let $G \in E$ be a fixed and publicly known point. The receiver $B$ chooses $a$ randomly and publishes the key $aG$, while keeping $a$ itself secret. To transmit a message $m$ to $B$, user $A$ chooses a random integer $k$ and sends the pair of points $(kG, P_m + k(aG))$. To read the message, $B$ multiplies the first point in the pair by his secret $a$, and then subtracts the result from the second point in the pair. So, if a breaker can compute $a$ from public key $G$ and $aG$, he can decrypt any encryption sent to $B$ (refer to [3]).

### 3.2. The construction of a parallel adder over $GF(2^n)$

Over $GF(2^n)$, the additive operation on two numbers is just doing *XOR* on each bit, respectively, without any carry. For instance, $(1101) + (1001) = (0100)$. For every bit $x_j$, two distinct 15 base value sequences are designed. One represents the value zero for $x_j$ and the other represents the value one for $x_j$. For convenience, we assume that $x^1_j$ denotes the value of $x_j$ to be one and $x^0_j$ denotes the value of $x_j$ to be zero. The following algorithm is used for parallel adding two $n$ bits binary number in a strand with one starts from the $p$th bit and the other one starts from the $q$th bit and appending the result from $r$th bit (see Algorithm 1).

Consider that $n = 3$, $p = 1$, $q = 4$, and $r = 7$. That is, two binary numbers to be added in parallel are both 3 bits, while one from the 1st bit and the other one from the 4th bit in a strand, and "append" operation starts from the 7th bit. We then suppose, tube $T_0 = \{110001, 010101, 001111, 100011\}$ which is regarded as an input tube for the algorithm ParallelAdder $(T_0, n, p, q, r)$. Because the value of $n$ is 3, step (1a) to step (1f) will be run 3 times. After the first execution of step (1a) is finished, $T_1 = \{110001, 100011\}$ and $T_2 = \{010101, 001111\}$. Next, after the first execution of step (1b) and step (1c) is performed, $T3 = \Phi$, $T_4 = \{110001, 100011\}$, $T_5 = \{010101, 001111\}$, and $T_6 = \Phi$, while $T_0 = T_1 = T_2 = \Phi$. After first execution of step (1d) is run, tube $T_7 = \{110001, 010101, 001111, 100011\}$ and $T8 = \Phi$. After first execution of step (1e) and step (1f) is run, $T_0 = \{1100011, 0101011, 0011111, 1000111\}$. Then, after the rest operations are performed, the result of tube $T_0$ is shown in Table 1.

**Lemma 1.** *The algorithm ParallelAdder $(T_0, n, p, q, r)$ is applied to finish the function of a parallel adder.*

Table 1: Result of tube $T_0$.

| Tube | The result is generated by ParallelAdder |
|---|---|
| $T_0$ | 110001111, 010101111, 001111110, 100011111 |

*Proof.* The algorithm ParallelAdder $(T_0, n, p, q, r)$ is implemented by means of the extract, merge, and append operations. Each execution of step (1a) is used to produce two tubes $T_1$ and $T_2$, where all of the molecules of DNA in $T_1$ contain $x^1_{p+j}$ and all of the molecules of DNA in $T_2$ contain $x^0_{p+j}$. Each execution of step (1b) and step (1c) is used to produce four tubes $T_3$, $T_4$, $T_5$, $T_6$, where all DNA strands in $T_3$ contain $x^1_{p+j}$ and $x^1_{q+j}$, all DNA strands in $T_4$ contain $x^1_{p+j}$ and $x^0_{q+j}$, all DNA strands in $T_5$ contain $x^0_{p+j}$ and $x^1_{q+j}$, and all DNA strands in $T_6$ contain $x^0_{p+j}$ and $x^0_{q+j}$. According to the additive theorem over $GF(2^n)$, the $j$th bit of the sum in $T_4$ and $T_5$ is 1 and the $j$th bit of the sum in $T_3$ and $T_6$ is 0.

From ParallelAdder $(T_0, n, p, q, r)$, it takes $3n$ extract operations, $3n$ merge operations, $2n$ append operations, and 9 test tubes to finish parallel addition. A value sequence for every bit contains 15 bases. Therefore, the algorithm will add $15n$ bases to all DNA strands in tube $T_0$. □

### 3.3. The construction of a parallel multiplier over $GF(2^n)$

Over $GF(2^n)$, the multiplicative operation runs as follows:

$$
\begin{aligned}
&(b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0)(b'_{n-1}\omega^{n-1} + \cdots + b'_1\omega + b'_0) \\
&\quad = b_{n-1}\omega^{n-1}(b'_{n-1}\omega^{n-1} + \cdots + b'_1\omega + b'_0) \\
&\qquad + \cdots + b_0(b'_{n-1}\omega^{n-1} + \cdots + b'_1\omega + b'_0) \\
&\quad = h_{2n-2}\omega^{2n-2} + h_{2n-3}\omega^{2n-3} + \cdots + h_2\omega^2 + h_1\omega + h_0,
\end{aligned}
$$

$$
\begin{aligned}
h_{2n-2} &= b_{n-1}b'_{n-1}, \\
h_{2n-3} &= b_{n-1}b'_{n-2} + b_{n-2}b'_{n-1}, \ldots, \\
h_{n-1} &= b_{n-1}b'_0 + b_{n-2}b'_1 + \cdots + b_0b'_{n-1}, \ldots, \\
h_1 &= b_1b'_0 + b_0b'_1, \\
h_0 &= b_0b'_0.
\end{aligned}
\tag{3}
$$

The algorithm ParallelMultiplier $(T_0, n, p, q)$ is used to multiply two $n$bit binary numbers on every strand in parallel with one starts from the $p$th bit and the other one starts from the $q$th bit. It runs as follows: at first, employ extract operation to form two tubes: $T_1$ and $T_2$. The first tube $T_1$ includes all of the strands on which $x_p = 1$ and the second tube $T_2$ includes all of the strands on which $x_p = 0$. Then, we copy the bits from $q$th to $(q + n - 1)$th to the end of every strand in tube $T_1$ and append $n$ bits 0 to the end of every strand in tube $T_2$. After these operations, the $(q + n)$th bit to $(q + n + n - 1)$th bit show the coefficients of $\omega^{2n-2}$ to $\omega^{n-1}$. Using the same principle, we get the coefficients of $\omega^{2n-3}$ to $\omega^{n-2}$, $\omega^{2n-4}$ to $\omega^{n-3}, \ldots, \omega^{n-1}$ to $\omega^0$. At last, call algorithm ParallelAdder $(T_0, n, p, q, r)$ to compute the sum of coefficients of $\omega^{2n-2}$, the sum of coefficient of $\omega^{2n-3}, \ldots$, and so on. As a result, the length of every strand will increase $n \times n + 2n - 1$ bits.

From ParallelMultiplier $(T_0, n, p, q)$, it takes $O(n^2)$ extract operations, $O(n^2)$ merge operations, $O(n^2)$ append operations, and $O(1)$ test tubes to finish the function of a parallel multiplier.

### 3.4. *The construction of a parallel shifter for multiplicative result*

Because over $GF(2^n)$, the exponent of $\omega$ cannot be beyond $n - 1$, the result of parallel multiplying should be shifted by a certain irreducible polynomial, called primitive polynomial: $\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0 = 0 \Leftrightarrow \omega^n = b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0$. The algorithm ParallelShifter $(T_0, n, p)$, $p$ representing that the multiplicative result starts from the $p$th bit, is used to parallel shift the multiplicative result $b_{2n-2}\omega^{2n-2} + b_{2n-3}\omega^{2n-3} + \cdots + b_1\omega + b_0$ to legal form $b_{n-1}\omega^{n-1} + b_{n-2}\omega^{n-2} + \cdots + b_1\omega + b_0$ over $GF(2^n)$ which can be designed as follows: appends the primitive polynomial's coefficients from $\omega^{n-1}$ to $\omega^0$ to the end of every strand at first. Employ the extract operation to form two test tubes $T_1$ and $T_2$. Tube $T_1$ includes all of the strands on which $x_p = 1$ and tube $T_2$ includes all of the strands on which $x_p = 0$. Then, we add the coefficients, from item $\omega^{2n-3}$ to item $\omega^{n-2}$, to the coefficients of irreducible polynomial in parallel in tube $T_1$. This forms the new coefficients from $\omega^{2n-3}$ to $\omega^{n-2}$ and has deleted the $\omega^{2n-2}$ item. The coefficients from $\omega^{n-3}$ to $\omega^0$ are without any change. For the $T_2$ includes all of the strands that have $x_p = 0$, so just copy the coefficients from $\omega^{2n-3}$ to $\omega^0$ without any change. After all the executions before are run, the highest exponent of $\omega$ is reduced to $2n - 3$. Then, merge $T_1$ and $T_2$ and begin new reduction. The principle of rest reducing turn is all like this above. When this algorithm is run out, the highest exponent of $\omega$ is reduced to $n - 1$. This algorithm will totally append $n \times n + (n - 1)(n - 2)/2$ bits more to every strand.

From ParallelShifter $(T_0, n, p)$, it takes $O(n^2)$ extract operations, $O(n^2)$ merge operations, $O(n^2)$ append operations, and $O(1)$ test tubes to finish the function of a parallel shifter.

### 3.5. *The mathematical principle of division on* $GF(2^n)$

Over $GF(2^n)$, to do a division operation for a dividend and a divisor, one should get the divisor's inverse first and then multiply the dividend. For the primitive polynomial $\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0$ is irreducible, there exists a polynomial $g(\omega)$ and a polynomial $f(\omega)$ that fit the equation (according to Euclid algorithm):

$$g(\omega) \times \text{divisor} + f(\omega) \times (\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0) = 1. \tag{4}$$

Because $\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0 = 0$, $g(\omega) \times \text{divisor} = 1$, which is to say $g(\omega)$ is the inverse of the divisor. To find $g(\omega)$ and $f(\omega)$, one can do as follows, which is called Euclid algorithm, also called division algorithm; first, $\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0$ is divided by the divisor. If the value of the remainder is 1, that is to say, $(\omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0) + g(\omega) \times \text{divisor} = 1$ with $g(\omega)$ is the division result and the inverse of divisor has found which is $g(\omega)$; else let remainder be $r(\omega)$, let the divisor be the dividend and $r(\omega)$ be the divisor and do division operation again. Repeat the process until the remainder is 1. Because the highest exponent of $\omega$ of remainder reduces by 1 in each repeat, it is at most repeating $n - 1$ times. So in the first time division, the dividend is $n + 1$ bits and the divisor is $n$ bits and the remainder is at most $n - 1$ bits, and in the second time division the dividend is $n$ bits and the divisor is $n - 1$ bits and the remainder is at most $n - 2$ bits, ..., and in last time division the dividend is 3 bits and the divisor is 2 bits and the remainder is 1 bit. Then, trace back to get the $g(\omega)$.

```
(1) If $(q - p > k - j)$ then
  (1a) For $m = 1$ to $q - p - (k - j)$
    (1a1) $T_3 = +(T_0, x^1_{p+m-1})$
    (1a2) $T_1 = \cup(T_1, T_3)$
  EndFor
EndIf
(2) For $m = q - (k - j) - p$ to $q - p$
  (2a) $T_3 = +(T_0, x^1_{p+m})$
  (2b) $T_4 = -(T_3, x^1_{j+m-(q-p-k+j)})$
  (2c) $T_5 = -(T_0, x^1_{p+m})$
  (2d) $T_6 = +(T_5, x^1_{j+m-(q-p-k+j)})$
  (2e) $T_1 = \cup(T_1, T_4)$ and $T_2 = \cup(T_2, T_3, T_6)$ and
       $T_0 = \cup(T_0, T_5)$
EndFor
EndProcedure
```

Algorithm 2: Procedure ParallelComparator $(T_0, p, q, j, k, T_1, T_2)$.

### 3.6. *The construction of a parallel comparator*

Prior to each step of long division, comparison should be done first. Suppose the divisor is $n$ bits at that time. At first compare the first two bits of dividend with divisor to determine that addition operation between first two bits of dividend and divisor should be done or not, then compare the first three bits of result with divisor, ..., compare the first $n$ bits of result with divisor, the last time (finally), compare the last $n$ bits of result with divisor for this time the first bit of result is 0. The following algorithm is used to compare the divisor which is from $p$th bit to $q$th bit with the bits from $j$th bit to $k$th bit in parallel, and forms tube $T_1$ and $T_2$. $T_2$ contains all strands on which add execution will be done and $T_1$ contains all strands on which add execution will not be done (see Algorithm 2).

**Lemma 2.** *The algorithm Parallel Comparator* $(T_0, p, q, j, k, T_1, T_2)$ *is applied to finish the function of parallel comparator.*

*Proof.* If $q - p > k - j$, that means the bits of divisor is more than the bits of the result which are intended to compare this

```
(1) For $j = 1$ to $n - 1$
  (1a) ParallelComparator $(T_0, p, p + n - 1, q + (j - 1)(n + 2), q + (j - 1)(n + 2) + j, T_1, T_2)$
  (1b) Append $(T_1, x^0_{q+n+1+(j-1)(n+2)})$ and Append $(T_2, x^1_{q+n+1+(j-1)(n+2)})$
  (1c) For $k = 0$ to $n$
    (1c1) $T_3 = +(T_1, x^1_{q+(n+2)(j-1)+k})$ and $T_4 = -(T_1, x^1_{q+(n+2)(j-1)+k})$
    (1c2) Append $(T_3, x^1_{q+j(n+2)+k})$ and Append $(T_4, x^0_{q+j(n+2)+k})$
    (1c3) $T_1 = \cup(T_3, T_4)$
  EndFor
  (1d) ParallelAdder $(T_2, j + 1, q + (j - 1)(n + 2), p + n - j - 1, q + (n + 2)j)$
  (1e) For $k = n - j$ down to 1
    (1e1) $T_3 = +(T_2, x^1_{q+(n+2)j-k-1})$ and $T_4 = -(T_2, x^1_{q+(n+2)j-k-1})$
    (1e2) Append $(T_3, x^1_{q+(j+1)(n+2)-k-1})$ and Append $(T_4, x^0_{q+(j+1)(n+2)-k-1})$
    (1e3) $T_2 = \cup(T_3, T_4)$
  EndFor
  (1f) $T_0 = \cup(T_1, T_2)$
EndFor
(2) ParallelComparator $(T_0, p, p + n - 1, q + (n + 2)(n - 1) + 1, q + (n + 2)(n - 1) + n, T_1, T_2)$
(3) Append $(T_1, x^0_{q+(n+2)n-1})$ and Append $(T_2, x^1_{q+(n+2)n-1})$
(4) For $k = 0$ to $n$
  (4a) $T_3 = +(T_1, x^1_{q+(n+2)(n-1)+k})$ and $T_4 = -(T_1, x^1_{q+(n+2)(n-1)+k})$
  (4b) Append $(T_3, x^1_{q+(n+2)n+k})$ and Append $(T_4, x^0_{q+(n+2)n+k})$
  (4c) $T_1 = \cup(T_3, T_4)$
EndFor
(5) Append $(T_2, x^0_{q+(n+2)n})$
(6) ParallelAdder $(T_2, n, q + (n + 2)(n - 1) + 1, p, q + (n + 2)n + 1)$
(7) $T_0 = \cup(T_1, T_2)$
EndProcedure
```

Algorithm 3: Procedure SimilarDiv $(T_0, n, p, q)$.

time. Step (1) considers the excessive bits of divisor and if any one bit is 1, which means the divisor is “bigger” than the bits of the result which are intended to compare this time and pour the strands to $T_1$. Step (2) considers the rest of the bits of divisor and the bits of result which are intended to compare this time. $T_3$ contains all strands on which certain bit of divisor is 1 and corresponding bit of the result is 1; $T_4$ contains all strands on which certain bit of divisor is 1 and corresponding bit of the result is 0; $T_5$ contains all strands on which certain bit of divisor is 0 and corresponding bit of the result is 0; $T_6$ contains all strands on which certain bit of divisor is 0 and corresponding bit of the result is 1. So add execution can be done over strands in $T_3$ and $T_6$, which can not be done over strands in $T_4$, and strands in $T_5$ need more consideration.

From ParallelComparator $(T_0, p, q, j, k, T_1, T_2)$, it takes $O(n)$ extract operations, $O(n)$ merge operations, and $O(1)$ test tubes to finish the function of a parallel comparator. □

### 3.7. The construction of a parallel long division

Suppose the divisor is $n$ bits and is from $p$th bit and the dividend is $n + 1$ bits and is from the $q$th bit in each strand. To do the long division, first compare the divisor with first two bits of dividend using ParallelComparator to get the first bit of the result of division and note down the result of first time addition result. Then, compare the divisor with the first three bits of the addition result last time to get the second bit of the result of the long division, and note down the addition result. Finally, compare the divisor with the last $n$ bits of addition result last time to get the last bit of division result and the remainder (see Algorithm 3).

**Lemma 3.** *The algorithm SimilarDiv $(T_0, n, p, q)$ is applied to finish the function of parallel long division.*

*Proof.* Each execution of step (1) is to get each bit of the long division result. The rest part is to get the last bit of the long division result. Consider the first cycle of step (1), step (1a) compares the divisor with the first two bits of the dividend, and form two tubes $T_1$ and $T_2$ that $T_2$ contains all strands on which add execution can be done, contrarily the $T_1$. Step (1b) appends 0 to all strands in $T_1$ and appends 1 to all strands in $T_2$. This bit is the first bit of the division result. Step (1c) just finishes to append the dividend in tube $T_1$. Step (1d) adds the divisor and the first two bits of the dividend in $T_2$. Step (1e) copies the rest of the bits of the dividend in $T_2$. Step (1f) pours $T_1$ and $T_2$ together and finishes the first execution of step (1) to get the first bit of the division result and the first time addition result. The second execution of step (1) is to compare the divisor with the first three bits of the addition result last time and get the second bit of the division result. The principle of other cycles and the rest of the steps are similar to the principle above. The length of each strand will reach to $q + n + (n + 2)n$ bits when this algorithm is run out.

```
(1) ParallelMultiplier ($T_0, n, d, p$)
(2) ParallelShifter ($T_0, n, p + n + n^2$)
(3) ParallelAdder ($T_0, n, t, p + M, p + n + M$)
EndProcedure
```

Algorithm 4: Procedure Traceback ($T_0, t, d, p, n$).

From SimilarDiv ($T_0, n, p, q$), it takes $O(n^2)$ extract operations, $O(n^2)$ append operations, $O(n^2)$ merge operations, and $O(1)$ test tubes to finish the function of a parallel long division. □

### 3.8. The construction of parallel traceback

Suppose the irreducible polynomial $A(\omega) = \omega^n + b_{n-1}\omega^{n-1} + \cdots + b_1\omega + b_0$ and suppose $g(\omega)$ and $f(\omega)$ satisfy that $g(\omega) \times$ divisor $+ f(\omega)A(\omega) = 1$, for the purpose of finding the divisor's inverse, $g(\omega)$, we need to do sometimes long division introduced in Section 3.7; let $A(\omega)$ be the dividend and divisor mentioned above be the divisor in first time and suppose the result is $g_1(\omega)$, and if the remainder is 1, then the division result $g_1(\omega)$ is $g(\omega)$. Else, let the divisor last time be the dividend and let the remainder last time be the divisor and do the long division. Suppose the result is $p(\omega)$ and $g_2(\omega) = p(\omega) \times g_1(\omega) + 1$, if the remainder is 1, the $g_2(\omega)$ is $g(\omega)$. Else, let the divisor last time be the dividend and let the remainder last time be the divisor and do the long division. Suppose the result is $p(\omega)$ and $g_3(\omega) = p(\omega) \times g_2(\omega) + g_1(\omega)$, if the remainder is 1, the $g_3(\omega)$ is $g(\omega)$. Generally speaking, we need to trace back after long division each time: first time, the tracing result is the division's result; the second time, the tracing's result is the sum of 1 and the product of the division's result and the tracing's result last time; from the third time, the tracing's result is the sum of the tracing's result last second time and the product of the division's result and the tracing's result last time. The following algorithm is used to do tracing operation from the third time in which $t$, $d$, and $p$ mean that the tracing's result last second time is from the $t$th bit and the last tracing's result is from the $d$th bit and the division result is from the $p$th bit (see Algorithm 4).

**Lemma 4.** *The algorithm TraceBack ($T_0, t, d, p, n$) is used to trace back after long division from the third time in order to get the divisor's inverse.*

*Proof.* In this and following procedures, $M = n^2 + 2n - 1 + n^2 + (n-1)(n-2)/2$, which represent the total number of increased bits when ParallelMultiplier ($T_0, n, p, q$) and ParallelShifter ($T_0, n, p$) are called. Step (1) is used to multiply the last tracing's result from $d$th bit to the long division result from $p$th bit in parallel. The result will be from the $(p + n + n \times n)$th bit to $(p + n + n \times n + 2n - 1)$th bit. Step (2) is to shift the result of multiplication to legal form which will append $((n-1)(n-2)/2 + n^2)$ bits to every strand. And its result is from $(p + M)$th bit to $(p + n + M - 1)$th bit. Step (3) is used to add the result to the last second time's tracing's result in parallel.

From TraceBack ($T_0, t, d, p, n$), it takes $O(n^2)$ extract operations, $O(n^2)$ append operations, $O(n^2)$ merge operations, and $O(1)$ test tubes to finish the function of tracing back. □

### 3.9. The construction of a parallel inverse

From the algorithms introduced above, we can find divisors' inverses in parallel as follows: first pick out the strands on which divisor equals to 1. Then, let the primitive polynomial be the dividend and the divisor be the divisor and do long division. Trace back to get the tracing's result first time and pick up the strands on which the remainder equals to 1 and store them in tube $T_1$. Then, let the divisor last time be the dividend and the remainder last time be the divisor and do long division. Collect the quotient and trace back. Pick up the strands on which the remainder equals to 1 and store them in tube $T_2, \ldots$, these executions, including long division, collecting every bit of the quotient and tracing back, are run $n - 1$ times at most. In the following algorithm, the parameters $n$ and $p$ mean that the divisor is $n$ bits and it starts from $p$th bit in every strand. The last parameter $r$ is used to represent that each strand in tube is $r - 1$ bit long and we begin append operation from $r$th bit of every strand.

Among the algorithm, the procedure Picking ($T_0, n, p, T_s$) is used to pick out the strands on which the $p$th bit to the $(p + n - 2)$th bit are all 0 and the $(p + n - 1)$th bit is 1 and store them in $T_s$. It is easy to program, so just omitted here (see Algorithm 5).

**Lemma 5.** *The algorithm ParallelInverse ($T_0, n, p, r$) is applied to find inverses over GF($2^n$) in parallel.*

*Proof.* Step (2) is used to append $n + 1$ bits irreducible polynomial, which is the dividend in first long division, to every strand. The execution of step (3) calls the algorithm SimilarDiv ($T_0, n, p, r$) to finish long division. Now the length of strands is added up to $r + n + (n + 2)n$ bits. Step (4) finishes the function of collecting every bit of quotient and will add $n$ bits to every strand. This is the first time tracing result. Step (5) calls the procedure Picking ($T_0, n, p, T_s$) to pick out the strands on which the remainder is 1 and store them in tube $T_1$. Step (6) finishes the operation of appending the dividend, which is divisor last time, to the strands. Note that $s_1 = n+1+(n+2)n+n$. Step (7) calls the algorithm SimilarDiv ($T_0, n, p, q$) to accomplish the second time long division. Step (8) employs the append operation to append a bit 0 in order to make the quotient to be $n$ bits. Step (9) finishes the function of collecting every bit of quotient and will add $n - 1$ bits to every strand. Steps (10) and (11) call the algorithm ParallelMultiplier ($T_0, n, p, q$) and ParallelShifter ($T_0, n, p$). Step (12) is used to add 1 to the product. These three steps accomplish the function of tracing back of the second time. Now the length of every strand is $r+s_1+n+(n+1)(n-1)+n+M+n-1$. Step (13) calls the algorithm Picking ($T_0, n, p, T_s$) to pick out the strands on which the remainder is 1 and store them in tube $T_2$. One execution of step (14) finishes the function of long division and tracing back and picking out the strands on which the remainder is 1. The step will be looped $n - 3$ times, because the long division should be done $n - 1$ times to make

```
(1) Picking $(T_0, n, p, T_z)$
(2) Append $A(x)$ to the end of all strands in $T_0$
(3) SimilarDiv $(T_0, n, p, r)$
(4) For $j = 0$ to $n - 1$
  (4a) $T_m = +(T_0, x^1_{r+n+1+(n+2)j})$ and $T_s = -(T_0, x^1_{r+n+1+(n+2)j})$
  (4b) Append $(T_m, x^1_{r+n+1+(n+2)n+j})$ and Append $(T_s, x^0_{r+n+1+(n+2)n+j})$
  (4c) $T_0 = \cup(T_m, T_s)$
EndFor
(5) Picking $(T_0, n - 1, r + n + 1 + (n + 2)(n - 1) + 3, T_1)$
(6) For $j = 0$ to $n - 1$
  (6a) $T_m = +(T_0, x^1_{p+j})$ and $T_s = -(T_0, x^1_{p+j})$
  (6b) Append $(T_m, x^1_{r+s_1+j})$ and Append $(T_s, x^0_{r+s_1+j})$
  (6c) $T_0 = \cup(T_m, T_s)$
EndFor
(7) SimilarDiv $(T_0, n - 1, r + n + 1 + (n + 2)(n - 1) + 3, r + s_1)$
(8) Append $(T_0, x^0_{r+s_1+n+(n+1)(n-1)})$
(9) For $j = 0$ to $n - 2$
  (9a) $T_m = +(T_0, x^1_{r+s_1+n+(n+1)j})$ and $T_s = -(T_0, x^1_{r+s_1+n+(n+1)j})$
  (9b) Append $(T_m, x^1_{r+s_1+n+(n+1)(n-1)+1+j})$ and Append $(T_s, x^0_{r+s_1+n+(n+1)(n-1)+1+j})$
  (9c) $T_0 = \cup(T_m, T_s)$
EndFor
(10) ParallelMultiplier $(T_0, n, r + n + 1 + (n + 2)n, r + s_1 + n + (n + 1)(n - 1))$
(11) ParallelShifter $(T_0, n, r + s_1 + n + (n + 1)(n - 1) + n + n^2)$
(12) Add 1 to the product above which will result in $n$ bits more to each strand
(13) Picking $(T_0, n - 2, r + s_1 + n + (n + 1)(n - 2) + 3, T_2)$
(14) For $j = 2$ to $n - 2$
  (14a) Copy dividend (divisor last time) to the end
  (14b) SimilarDiv $(T_0, n - j, r + s_1 + \cdots + s_{j-1} + n + 2 - j + (n + 3 - j)(n - j) + 3, r + s_1 + \cdots + s_j)$
  (14c) Append $j$ bits 0 to the end
  (14d) Collect $n - j$ bits quotient of this division to the end
  (14e) Traceback $(T_0, r + s_1 + \cdots + s_{j-1} - n, r + s_1 + \cdots + s_j - n, r + s_1 + \cdots + s_j + (n + 1 - j) + (n + 2 - j)(n - j), n)$
  (14f) Picking $(T_0, n - 1 - j, r + s_1 + \cdots + s_j + n + 1 - j + (n + 2 - j)(n - 1 - j) + 3, T_{j+1})$
EndFor
(15) For $j = 1$ to $n - 1$
  (15a) For $k = 0$ to $s_{j+1} + \cdots + s_{n-1} - 1$
    (15a1) Append $(T_j, x^0_{r+s_1+\cdots+s_j+k})$
  EndFor
  (15b) For $k = 0$ to $n - 1$
    (15b1) $T_m = +(T_j, x^1_{r+s_1+\cdots+s_j-n+k})$ and $T_s = -(T_j, x^1_{r+s_1+\cdots+s_j-n+k})$
    (15b2) Append $(T_m, x^1_{r+s_1+\cdots+s_{n-1}+k})$ and Append $(T_s, x^0_{r+s_1+\cdots+s_{n-1}+k})$
    (15b3) $T_j = \cup(T_m, T_s)$
  EndFor
EndFor
(16) For $j = 0$ to $s_1 + \cdots + s_{n-1} + n - 2$
  Append $(T_z, x^0_{r+j})$
EndFor
(17) Append $(T_z, x^1_{r+s_1+\cdots+s_{n-1}+n-1})$
(18) $T_0 = \cup(T_z, T_1, T_2, \ldots, T_{n-1})$
EndProcedure
```

Algorithm 5: Procedure ParallelInverse $(T_0, n, p, r)$.

sure that the remainder of each strand equals to 1 and long division has been done twice before. Note that while $j \geq 2$, $s_j = n + 2 - j + (n + 3 - j)(n + 1 - j) + n + M + n$. Step (15a) is used to append certain bits 0 to each tube $T_j$ to make sure that strands in $T_1$ to $T_{n-1}$ are all $r + s_1 + \cdots + s_{n-1} - 1$ bits long. Step (15b) is used to append the inverse gotten before to the last of every strand in $T_1$ to $T_{n-1}$. Step (18) pours strands in all tubes together to one tube $T_0$. From all above, getting inverse one time needs increasing $s_1 + s_2 + \cdots + s_{n-1} + n$ bits to every strand.

From ParallelInverse $(T_0, n, p, r)$, it takes $O(n^3)$ extract operations, $O(n^3)$ append operations, $O(n^3)$ merge operations, and $O(n)$ test tubes to finish the function of finding inverse in parallel. □

```
(1) ParallelInverse $(T_0, n, p, r)$
(2) ParallelMultiplier $(T_0, n, q, r + s_1 + \cdots + s_{n-1})$
(3) ParallelShifter $(T_0, n, r + s_1 + \cdots + s_{n-1} + n + n \times n)$
EndProcedure
```

Algorithm 6: Procedure ParallelDivision $(T_0, n, p, q, r)$.

### 3.10. *The construction of a parallel divider*

To do a division operation on $GF(2^n)$, first one should calculate divisor's inverse using above mentioned algorithm then multiply the dividend. The following algorithm is used to finish the function of parallel division over $GF(2^n)$, where the dividend begins with the $q$th bit and the divisor begins with the $p$th bit, and current bit is the $r$th bit (see Algorithm 6).

**Lemma 6.** *The algorithm ParallelDivision $(T_0, n, p, q, r)$ is used to finish the function of parallel division over $GF(2^n)$.*

*Proof.* Step (1) calls the algorithm ParallelInverse $(T_0, n, p, r)$ to get the divisor's inverse of every strand. The length of every strand is $r+s_1+\cdots+s_{n-1}+n-1$ bits now with the inverse from $(r+s_1+\cdots+s_{n-1})$th bit to $(r+s_1+\cdots+s_{n-1}+n-1)$th bit. Step (2) calls ParallelMultiplier $(T_0, n, p, q)$ to finish the function of the inverse being multiplied by the dividend every strand with the dividend starting from the $q$th bit. Now the length of each strand adds up to $r+s_1+\cdots+s_{n-1}+n+n\times n+2n-1-1$ bits. Step (3) shifts the product by calling ParallelShifter $(T_0, n, p)$ and will add $(n-1)(n-2)/2 + n \times n$ bits to every strand. Generally speaking, doing parallel division one time need increasing the strands $D = s_1 + \cdots + s_{n-1} + n + M$ bits.

From ParallelDivision $(T_0, n, p, q, r)$, it takes $O(n^3)$ extract operations, $O(n^3)$ append operations, $O(n^3)$ merge operations, and $O(n)$ test tubes to finish the function of parallel division. □

### 3.11. *The construction of a parallel adder of two points on elliptic curve*

By far the addition, subtraction (the same to addition), multiplication, and division operations over $GF(2^n)$ have been solved. Now consider how to execute addition of two points on an elliptic curve $y^2 + xy = x^3 + ax^2 + b$ in biological ways. We should consider five different cases: case 1, the first point is $O$ and the point of sum equals to the second point; case 2, the second point is $O$ and the point of sum equals to the first point; case 3, one point is the inverse of the other one, and the point of sum is $O$; case 4, one point equals to the other one, and computes the sum as the formula in part 3.1; case 5, computes the sum using the formula in part 3.1.

The algorithm AddTwoNode $(T_0, n, x1, y1, x2, y2, r)$ is used to add two points. The first one's position $x$ starts from the $(x1)$th bit and position $y$ starts from the $(y1)$th bit, and the second one's position $x$ starts from the $(x2)$th bit and $y$ starts from the $(y2)$th bit. The parameter $r$ represents the current bit. In the procedure, it calls Picking01 $(T_0, n, x1, y1, x2, y2, T_{11})$ to pick out the strands on which the first point is $O$ and store them in $T_{11}$, Picking02 $(T_0, n, x1, y1, x2, y2, T_{12})$ to pick out the strands on which the second point is $O$ and store them in $T_{12}$, PickingInverse $(T_0, n, x1, y1, x2, y2, r, T_2)$ to pick out the strands on which one point is the inverse of the other one and store them in tube $T_2$, and PickingEqual $(T_0, n, x1, y1, x2, y2, T_3)$ to pick out the strands on which one point equals to the other one and store them in tube $T_3$. These four algorithms are easy to design, so omitted here. Note that PickingInverse $(T_0, n, x1, y1, x2, y2, r, T_2)$ will increase $n$ bits to every strand in $T_0$ (see Algorithm 7).

**Lemma 7.** *The algorithm AddTwoNode $(T_0, n, x1, y1, x2, y2, r)$ can compute the sum of two points on elliptic curve.*

*Proof.* Step (5) employs append operation to append $M$ bits 0 to all strands in tube $T_0$. Step (6) to step (8) are operations on tube $T_0$ to get $\mu$ of strands in $T_0$. Step (9) to step (13) are operations on tube $T_3$ to get $\mu$ of strands in $T_3$. Step (14) pours $T_0$ and $T_3$ to $T_0$ and the length of every strand in $T_0$ now is $r + n + n + M + n + D - 1$ bits. Step (15) to step (21) accomplish to compute the position $x$ of point of sum and the result is from $(r+X)$th bit. The execution of steps (22) to (26) is used to get the position $y$ of the sum which equals to $\mu(x1 + x3) + x3 + y1$. And now, the position $y$ of the sum is from $(r+X+Y)$th bit and the length is $r+n+X+Y-1$ bits.

Steps (27) and (28) are applied to append the value of positions $x$ and $y$ of the sum of two points to the last of every strand. Now, the length of every strand in $T_0$ is $r + n + X + Y + 2n - 1$ bits with the value $x$ from the $(r + n + X + Y)$th bit and $y$ from the $(r + n + X + Y + n)$th bit.

From AddTwoNode $(T_0, n, x1, y1, x2, y2, r)$, it takes $O(n^3)$ extract operations, $O(n^3)$ append operations, $O(n^3)$ merge operations, and $O(n)$ test tubes to finish the function of a parallel adder for points on elliptic curve. □

### 3.12. *Breaking the elliptic curve cryptosystem*

We have constructed the algorithm above for parallel computing the point of the sum of two points. Then, we can solve elliptic curve discrete logarithm problem as follows: consider point $P$ and $Q$ are given, and $l$ is what we want to get which matches $Q = lP$. First, we amplify $P$ into two tubes and add $P$ in one tube. Check if $2P$ equals to $Q$; if not, note down the value of $2P$ and pour two tubes together. Then, amplify the tube into two tubes and add $2P$ in one tube. Check if any point equals to $Q$; if not, note down the value of $4P$ and pour two tubes together, or we get the value of $l$. Then, amplify the tube into two tubes and add $4P$ in one tube, . . . , while this loop executes $n$ times, the value from 1 to $2^n$ for $l$ will have been checked, and the elliptic curve cryptosystem has been broken by the solved elliptic curve discrete logarithm problem.

## 4. CONCLUSION

This paper is the first effort in literature that demonstrates that the difficult problem for elliptic curve discrete logarithm can be solved on a DNA-based computer. While Chang's

```
(1) Picking01 $(T_0, n, x1, y1, x2, y2, T_{11})$
(2) Picking02 $(T_0, n, x1, y1, x2, y2, T_{12})$
(3) PickingInverse $(T_0, n, x1, y1, x2, y2, r, T_2)$
(4) PickingEqual $(T_0, n, x1, y1, x2, y2, T_3)$
(5) For $j = 0$ to $M - 1$
    (5a) Append $(T_0, x^0_{r+n+j})$
EndFor
(6) ParallelAdder $(T_0, n, y1, y2, r + n + M)$
(7) ParallelAdder $(T_0, n, x1, x2, r + n + M + n)$
(8) ParallelDivision $(T_0, n, r + n + M + n, r + n + M, r + n + M + 2n)$
(9) For $j = 0$ to $n - 1$
    (9a) $T_m = +(T_3, x^1_{x1+j})$ and $T_s = -(T_3, x^1_{x1+j})$
    (9b) Append $(T_m, x^1_{r+n+j})$ and Append $(T_s, x^0_{r+n+j})$
    (9c) $T_3 = \cup(T_m, T_s)$
EndFor
(10) ParallelMultiplier $(T_3, n, r + n, r + n)$
(11) ParallelShifter $(T_3, n, r + n + n + n \times n)$
(12) ParallelAdder $(T_3, n, y1, r + n + n + M - n, r + n + n + M)$
(13) ParallelDivision $(T_3, n, x1, r + n + n + M, r + n + n + M + n)$
(14) $T_0 = \cup(T_0, T_3)$
SUPPOSE $U = n + M + n + D$
(15) ParallelMultiplier $(T_0, n, r + n + U - n, r + n + U - n)$
(16) ParallelShifter $(T_0, n, r + n + U + n \times n)$
(17) ParallelAdder $(T_0, n, r + n + U - n, r + n + U + M - n, r + n + U + M)$
(18) ParallelAdder $(T_0, n, r + n + U + M, x1, r + n + U + M + n)$
(19) ParallelAdder $(T_0, n, r + n + U + M + n, x2, r + n + U + M + n + n)$
(20) For $j = 0$ to $n - 1$
    (20a) Append $(T_0, x_{a,r+n+U+M+3n+j})$
EndFor
(21) ParallelAdder $(T_0, n, r + n + U + M + 2n, r + n + U + M + 3n, r + n + U + M + 4n)$
SUPPOSE $X = U + M + 5n$
(22) ParallelAdder $(T_0, n, x1, r + n + X - n, r + n + X)$
(23) ParallelMultiplier $(T_0, n, r + n + U - n, r + n + X)$
(24) ParallelShifter $(T_0, n, r + n + X + n + n \times n)$
(25) ParallelAdder $(T_0, n, r + n + X + n + M - n, r + n + X - n, r + n + X + n + M)$
(26) ParallelAdder $(T_0, n, r + n + X + n + M, y1, r + n + X + n + M + n)$
SUPPOSE $Y = n + M + n + n$
(27) For $j = 0$ to $n - 1$
    (27a) $T_m = +(T_0, x^1_{r+X+j})$ and $T_s = -(T_0, x^1_{r+X+j})$
    (27b) Append $(T_m, x^1_{r+n+X+Y+j})$ and Append $(T_s, x^0_{r+n+X+Y+j})$
    (27c) $T_0 = \cup(T_m, T_s)$
EndFor
(28) For $j = 0$ to $n - 1$
    (28a) $T_m = +(T_0, x^1_{r+n+X+Y-n+j})$ and $T_s = -(T_0, x^1_{r+n+X+Y-n+j})$
    (28b) Append $(T_m, x^1_{r+n+X+Y+n+j})$ and Append $(T_s, x^0_{r+n+X+Y+n+j})$
    (28c) $T_0 = \cup(T_m, T_s)$
EndFor
(29) Append $n + X + Y$ bits 0 to each strand in $T_{11}$ and $T_{12}$. Then, append values of $x2$ and $y2$ of each strand to
        the end in $T_{11}$, and append values of $x1$ and $y1$ of each strand to the end in $T_{12}$
(30) Append $X + Y + 2n$ bits 0 to every strand in $T_2$
(31) $T_0 = \cup(T_{11}, T_{12}, T_2, T_0)$
EndProcedure
```

Algorithm 7: Procedure AddTwoNode $(T_0, n, x1, y1, x2, y2, r)$.

work makes great progress in application of DNA computing in cryptoanalysis [16], which is breaking RSA by factoring integer, this paper proposes application of DNA computing in another popular cryptosystem, ECC, which is more complex and has more challenge in cryptoanalysis. Though the algorithm is somewhat complex, it takes a series of steps that is polynomial in the input size, so it is feasible in theory and inspirits the development of DNA computing. Simultaneously, the paper also shows that humans' complex mathematical operations can directly be performed with basic biological operations. The property for the difficulty of elliptic curve discrete logarithm is the basis of elliptic curve cryptosystems. However, this property seems to be incorrect on a molecular computer. This indicates that the elliptic curve cryptosystems are perhaps insecure if the technique of DNA computing is skillful enough to run the algorithms efficiently as discussed in this paper.

## ACKNOWLEDGMENTS

The authors thank Daniel Howard (QinetiQ, Malvern Technology Centre, UK) for his great efforts in correcting this paper. This work is supported by National Natural Science Foundation (60603053) of China.

## REFERENCES

[1] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 5187, pp. 1021–1024, 1994.

[2] E. Csuhaj-Varjú, L. Kari, and G. Paun, "Test tube distributed systems based on splicing," *Computers and Artificial Intelligence*, vol. 15, no. 2-3, pp. 211–232, 1996.

[3] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of Computation*, vol. 48, no. 177, pp. 203–209, 1987.

[4] N. Koblizt, *Introduction to Elliptic Curves and Modular Forms*, Springer, New York, NY, USA, 1984.

[5] S. Lang, *Elliptic Curves: Diophantine Analysis*, Springer, New York, NY, USA, 1978.

[6] V. S. Miller, "Use of elliptic curves in cryptography," in *Proceedings of the 5th Annual International Cryptology Conference (CRYPTO '85)*, Santa Barbara, Calif, USA, August 1985.

[7] A. M. Odlyzko, "Discrete logarithms in finite fields and their cryptographic significance," in *Proceedings of the 2nd Workshop on Advances in Cryptology: Theory and Application of Cryptographic Techniques (EUROCRYPT '84)*, pp. 224–314, Springer, Paris, France, April 1985.

[8] R. P. Feynman, "There's plenty of room at the bottom," in *Minaturization*, D. H. Gilbert, Ed., pp. 282–296, Reinhold, New York, NY, USA, 1961.

[9] R. J. Lipton, "DNA solution of hard computational problems," *Science*, vol. 268, no. 5210, pp. 542–545, 1995.

[10] R. R. Sinden, *DNA Structure and Function*, Academic Press, New York, NY, USA, 1994.

[11] G. Paun, G. Rozenberg, and A. Salomaa, *DNA Computing: New Computing Paradigms*, Springer, New York, NY, USA, 1998.

[12] W.-L. Chang, M. S.-H. Ho, and M. Guo, "Molecular solutions for the subset-sum problem on DNA-based supercomputing," *Biosystems*, vol. 73, no. 2, pp. 117–130, 2004.

[13] M. Guo, M. S.-H. Ho, and W.-L. Chang, "Fast parallel molecular solution to the dominating-set problem on massively parallel bio-computing," *Parallel Computing*, vol. 30, no. 9-10, pp. 1109–1125, 2004.

[14] M. S.-H. Ho, "Fast parallel molecular solutions for DNA-based supercomputing: the subset-product problem," *Biosystems*, vol. 80, no. 3, pp. 233–250, 2005.

[15] D. Boneh, C. Dunworth, and R. J. Lipton, "Breaking DES using a molecular computer," Tech. Rep. CS-TR-489-95, Princeton University, Princeton, NJ, USA, 1995.

[16] W.-L. Chang, M. Guo, and M. S.-H. Ho, "Fast parallel molecular algorithms for DNA-based computation: factoring integers," *IEEE Transactions on Nanobioscience*, vol. 4, no. 2, pp. 149–163, 2005.

[17] J. Watson, M. Gilman, J. Witkowski, and M. Zoller, *Recombinant DNA*, Freeman, San Francisco, Calif, USA, 2nd edition, 1992.

[18] F. Eckstein, *Oligonucleotides and Anologues*, Oxford University Press, Oxford, UK, 1991.

[19] J. Watson, N. Hoplins, J. Roberts, J. Steitz, and A. Weiner, *Molecular Biology of the Gene*, Benjamin/Cummings, Menlo Park, Calif, USA, 1987.

[20] G. M. Blackburn and M. J. Gait, *Nucleic Acids in Chemistry and Biology*, IRL Press, Washington, DC, USA, 1990.

[21] M. Wiener and R. Zuccherato, "Faster attacks on elliptic curve cryptosystems," in *Selected Areas in Cryptography*, vol. 1556 of *Lecture Notes in Computer Science*, pp. 190–200, Springer, New York, NY, USA, 1999.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 218097, 7 pages
doi:10.1155/2008/218097

*Research Article*

# Classification Models for Early Detection of Prostate Cancer

**Joerg D. Wichard,[1, 2] Henning Cammann,[1] Carsten Stephan,[3] and Thomas Tolxdorff[1]**

[1] *Institute of Medical Informatics, Charité - Universitätsmedizin, Hindenburgdamm 30, 12200 Berlin, Germany*
[2] *Molecular Modelling Group, Institut für Molekulare Pharmakologie, Robert Rössle Straße 10, 13125 Berlin, Germany*
[3] *Department of Urology, Charité - Universitätsmedizin, Charitéplatz 1, 10098 Berlin, Germany*

Correspondence should be addressed to Joerg D. Wichard, joergwichard@web.de

Received 12 September 2007; Accepted 2 January 2008

Recommended by Daniel Howard

We investigate the performance of different classification models and their ability to recognize prostate cancer in an early stage. We build ensembles of classification models in order to increase the classification performance. We measure the performance of our models in an extensive cross-validation procedure and compare different classification models. The datasets come from clinical examinations and some of the classification models are already in use to support the urologists in their clinical work.

## 1. INTRODUCTION

Prostate cancer is one of the most common types of cancer among male patients in the western world. The number of expected new cases in the USA for the year 2006 was 235,000 with 27,000 expected deaths [1]. Early detection of prostate cancer improves the chances of a curative treatment and a lot of progress has been made in this field during the last decade. The early detection is considerably enhanced by the measurement of prostate-specific antigen (PSA) in conjunction with other clinically available data like age, digital rectal examination (DRE), and transrectal ultrasonography (TRUS) variables like prostate volume. We compared several classification models and analyzed their performance on the clinical dataset with an extended cross-validation procedure. The models were linear discriminant analysis (LDA), penalized discriminant analysis (PDA) [2], logistic regression [3], classification and regression trees (CARTs) [4], multilayer perceptron (MLP) [5], support vector machines (SVMs) [6, 7], and nearest neighbour classifiers [8]. All these models are implemented in an open-source Matlab-toolbox that is available on the internet [9].

This study will help to improve the software package *ProstataClass* [10] which was developed at Charité and currently uses an artificial neural network as classification engine. This program is successfully used in clinical practice for several years.

## 2. DATA

We had access to the clinically available data of 506 patients with 313 cases of prostate cancer (PCa) and 193 non-PCa. The data were selected from a group of 780 patients randomly. The data entry for each patient included age, PSA, the ratio of free to total prostate-specific antigen (PSA-Ratio), TRUS, and the diagnostic finding from the DRE which was a binary variable (suspicious or nonsuspicious). Blood sampling and handling were performed as described in Stephan et al. [11]. The samples were taken before any diagnostic or therapeutic procedures, and sera were stored at 80°C until analyzed. After thawing at room temperature, samples were processed within 3 hours. Prostate volume was determined by transrectal ultrasound using the prolate ellipse formula. The scatter plot of the variables under investigation is shown in Figure 1. PCa and non-PCa patients were histologically confirmed by 6–8 core prostate biopsy.

## 3. ENSEMBLES

The average output of several different models $f_i(x)$ is called an ensemble model:

$$\hat{f}(x) = \sum_{i=1}^{K} \omega_i f_i(x), \tag{1}$$

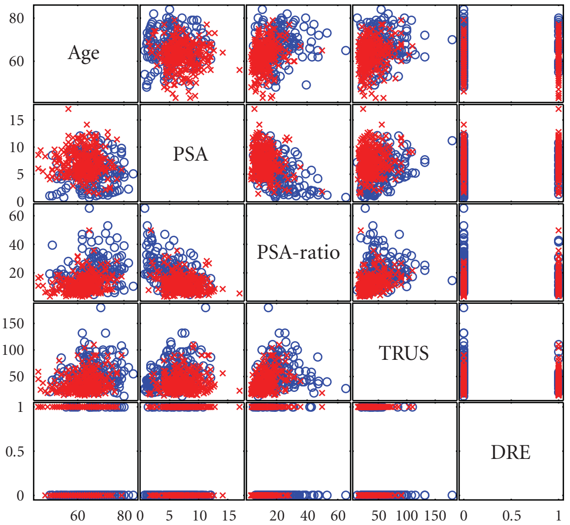


FIGURE 1: A scatterplot matrix of the data. Each box shows a pair of variables and the cases are color-coded, a red cross marks PCa, and a blue circle non-PCa. The DRE is a binary variable (suspicious or nonsuspicious).

where we assume that the model weights $\omega_i$ sum to one $\sum_{i=1}^{K}\omega_i = 1$. There are several suggestions concerning the choice of the model weights (see Perrone and Cooper [12]) but we decided to use uniform weights with $\omega_i = 1/K$ for the sake of simplicity and not to run into overfitting problems as reported by Krogh and Sollich [13].

The central feature of the ensemble approach is the generalization ability of the resulting model. In the case of regression models (with continuous output values), it was shown that the generalization error of the ensemble is in the average case lower than the mean of the generalization error of the single-ensemble members (see Krogh and Vedelsby 1995 [14]). This holds in general, independent of the model class, as long as the models constituting the ensemble are diverse with respect to the hypothesis of the unknown function. In the case of (binary) classification models, the situation was not so clear because the classical bias-variance decomposition of the squared error loss in regression problems (Geman et al. [15]) had to be extended to the zero-one loss function. There are several approaches dealing with this problem, see Kong and Dietterich [16], Kohavi and Wolpert [17], or Domingos [18].

The zero-one loss function is not the only possible choice for classification problems. If we are interested in a likelihood whether a sample belongs to one class or not, we can use the error loss from regression and consider the binary classification problem as a regression problem that works on two possible outcomes. In practice, many classifiers are trained in that way.

Our ensemble approach is based on the observation that the generalization error of an ensemble model could be improved if the models on which averaging is done disagree and if their residual errors are uncorrelated [13]. To see this, we have to investigate the contribution of the single model in the ensemble to the generalization error. We consider the case where we have a given dataset $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ and we want to find a function $f(\mathbf{x})$ that approximates $y$ at new observations of $\mathbf{x}$. These observations are assumed to come from the same source that generated the training set $D$, that is, from the same (unknown) probability distribution $P$. It should be noted that $f$ depends also on $D$. The expected generalization error $\mathrm{Err}(\mathbf{x}, D)$ given a particular $\mathbf{x}$ and a training set $D$ is

$$\mathrm{Err}(\mathbf{x}, D) = E\left[(y - f(\mathbf{x}))^2 \mid \mathbf{x}, D\right], \tag{2}$$

where the expectation $E[\cdot]$ is taken with respect to the probability distribution $P$. We are now interested in

$$\mathrm{Err}(\mathbf{x}) = E_D[\mathrm{Err}(\mathbf{x}, D)], \tag{3}$$

where the expectation $E_D[\cdot]$ is taken with respect to all possible realizations of training sets $D$ with fixed sample size $N$. According to Geman et al. [15], the bias/variance decomposition of $\mathrm{Err}(\mathbf{x})$ is

$$\begin{aligned}\mathrm{Err}(\mathbf{x}) &= \sigma^2 + (E_D[f(\mathbf{x})] - E[y \mid \mathbf{x}])^2 \\ &\quad + E_D\left[(f(\mathbf{x}) - E_D[f(\mathbf{x})])^2\right] \\ &= \sigma^2 + \mathrm{Bias}(f(x))^2 + \mathrm{Var}(f(x)),\end{aligned} \tag{4}$$

where $E[y \mid \mathbf{x}]$ is the deterministic part of the data and $\sigma^2$ is the variance of $y$ given $\mathbf{x}$. Balancing between the bias and the variance terms is a crucial problem in model building. If we try to decrease the bias term on a specific training set, we usually increase the variance term and vice versa. We now consider the case of an ensemble average $\hat{f}(\mathbf{x})$, consisting of $K$ individual models as defined in (1). If we put this into (4), we get

$$\mathrm{Err}(\mathbf{x}) = \sigma^2 + \mathrm{Bias}(\hat{f}(x))^2 + \mathrm{Var}(\hat{f}(x)), \tag{5}$$

and we can have a look at the effects concerning bias and variance. The bias term in (5) is just the average of the biases of the individual models in the ensemble. So we should not expect a reduction in the bias term compared to single models. According to Naftaly et al. [19], the variance term of the ensemble could be decomposed in the following way:

$$\begin{aligned}\mathrm{Var}(\hat{f}) &= E\left[\left(\hat{f} - E\left[\hat{f}\right]\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^{K}\omega_i f_i\right)^2\right] - \left(E\left[\sum_{i=1}^{K}\omega_i f_i\right]\right)^2 \\ &= \sum_{i=1}^{K}\omega_i^2 (E[f_i^2] - E^2[f_i]) \\ &\quad + 2\sum_{i<j}\omega_i\omega_j (E[f_i f_j] - E[f_i]E[f_j]),\end{aligned} \tag{6}$$

where the expectation is taken with respect to $D$. The first sum in (6) marks the lower bound of the ensemble variance and is the weighted mean of the variances of the ensemble members. The second sum contains the cross terms of the ensemble members and disappears if the models are completely uncorrelated [13]. So the reduction in the variance of the ensemble is related to the degree of independence of the single models [19].

## 4. CROSS-VALIDATION AND MODEL SELECTION

Our model selection scheme is a mixture of bagging [20] and cross-validation. *Bagging* or *Bootstrap aggregating* was proposed by Breiman [20] in order to improve the classification by combining classifiers trained on randomly generated subsets of the entire training sets. We extended this approach by applying a cross-validation scheme for model selection on each subset and after that we combine the selected models to an ensemble. In contrast to classical cross-validation, we use random subsets as cross-validation folds. In $K$-fold cross-validation, the dataset is partitioned into $K$ subsets. Of these $K$ subsets, a single subset is retained as the validation data for testing the model, and the remaining $K - 1$ subsets are used for model training. The cross-validation process is then repeated $K$ times with each of the $K$ subsets used only once as the validation data. The $K$ results from the folds then can be averaged to produce a single estimation.

If we lack relevant problem-specific knowledge, cross-validation methods could be used to select a classification method empirically [21]. This is a common approach because it seems to be obvious that no classification method is uniformly superior, see, for example, Quinlan [22] for a detailed study. It is also a common approach to select the model parameters with cross-validation [23]. The idea to combine the models from the $K$ cross-validation folds (stacking) was described by Wolpert [24].

We suggest to train several models on each CV-fold, to select the best performing model on the validation set, and to combine the selected models from the $K$-folds. If we train models of one type but with different initial conditions (e.g., neural networks with different numbers of hidden neurons), then we find proper values for the free parameters of the model. We could extend that by combining models from different classes in order to increase the model diversity. We call this a *heterogeneous ensemble* or *mixed ensemble* and applied this method effectively to regression problems [25] and classification tasks [26].

Our model selection scheme works as follows: for the $K$-fold CV, the data is divided $K$-times into a *training set* and a *test set*, both sets containing randomly drawn subsets of the data without replications. The size of each test set was 25% of the entire dataset.

In every CV-fold, we train several different models with a variety of model parameters (see Section 5 for an overview of the models and the related model parameters). In each fold, we select only one model to become a member of the final ensemble (namely, the best model with respect to the test set). This means that all models have to compete with each other in a fair tournament because they are trained and validated on the same dataset. The models with the lowest classification error in each CV-fold are taken out and added to the final ensemble, receiving the weight $\omega_i = 1/k$ (see (1)). All other models in this CV-fold are deleted.

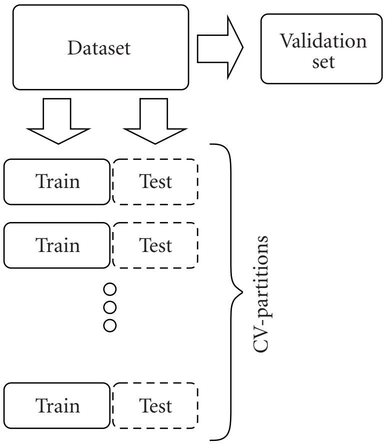


FIGURE 2: For every partition of the cross-validation, the data is divided in a training and a test set. The performance of each ensemble model was assessed on validation set which was initially removed and never included in model training.

We can use this model selection scheme in two ways. If we have no idea or prior knowledge, where classification or regression method should be used to cope with a specific problem, we could use this scheme in order to look for an empirical answer and to compare the performance of the different model classes. The other way is the estimation of model parameters for the different model classes described in Section 5.

## 5. CLASSIFICATION MODELS

In this section, we give a short overview of the model classes that we used for ensemble building. All models belong to the well-established collection of machine-learning algorithms for classification and regression tasks, so details can be found in the textbooks like, for instance, Hastie et al. [2]. The implementation of these models in an open-source toolbox together with a more detailed description can be found in [9]. The toolbox is an open-source MATLAB Toolbox which allows the integration of existing implementations of classification algorithms and it contains more then the few model classes described in the text.

### 5.1. *Linear discriminant analysis*

The LDA is a simple but useful classifier. If we assume that the two classes $k = \{0, 1\}$ have a Gaussian distribution with mean $\mu_k$ and they share the same covariance matrix $\Sigma$, then the *linear discriminant function* $\delta_k(\mathbf{x})$, $k = \{0, 1\}$ is given by

$$\delta_k(\mathbf{x}) = \mathbf{x}^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k), \qquad (7)$$

where $\pi_k$ denotes the frequency of occurrence of the class labels. The predicted class labels are given by

$$f(\mathbf{x}) = \arg\max_{k=(0,1)}\{\delta_k(\mathbf{x})\}. \qquad (8)$$

We also implemented two modifications: the quadratic discriminant analysis (QDA) and the PDA, as described in detail in Hastie et al. [2]. Linear method are usually conceptually simple, robust, fast, and, in particular in high-dimensional problems, they could be very powerful.

### 5.2. *Logistic regression model*

Logistic regression (Log.Reg.) is a model for binomial distributed dependent variables and is used extensively in the medical and social sciences. Hastie et al. [2] pointed out that the Logistic Regression model has the same form as the LDA, the only difference lies in the way, the linear coefficients are estimated. See Hosmer and Lemeshow for a detailed introduction [27]. We used the binary Log.Reg. to compute the probability of the dichotomic variable $y$ (PCa or non-PCa) from the $m$ independent variables $\mathbf{x}$:

$$y = \frac{1}{1+\exp(\mathbf{z})} \tag{9}$$

with

$$\mathbf{z} = a_0 + \sum_{i=1}^{m} a_i x_i, \tag{10}$$

wherein the model coefficients are estimated with a second-order gradient decent (quadratic approximation to likelihood function). This could be a critical issue in high-dimensional problems because these calculations are time and memory consuming.

### 5.3. *Multilayer perceptron*

We train a multilayer feed-forward neural network "MLP" with a sigmoid activation function. The weights are initialized with Gaussian-distributed random numbers having zero mean and scaled variances. The weights are trained with a gradient descend based on the Rprop algorithm [28] with the improvements given in [29]. The MLP works with a common weight decay with the penalty term

$$P(\vec{w}) = \lambda \sum_{i=1}^{N} - \frac{w_i^2}{1+w_i^2}, \tag{11}$$

where $\vec{w}$ denotes the $N$-dimensional weight vector of the MLP and a small regularization parameter $\lambda$. The number of hidden layers, the number of neurons, and the number of regularization parameter are adjusted during the CV-training. We further applied the concept of an $\epsilon$-insensitive error loss that we introduced in the context of cellular neural networks (CNNs) [30].

### 5.4. *Support vector machines*

Over the last decade, SVMs have become very powerful tools in machine learning. An SVM creates a hyperplane in a feature space that separates the data into two classes with the maximum margin. The feature space can be a mapping of the original features $(\mathbf{x}, \mathbf{x}')$ into a higher-dimensional space using a positive semidefinite function:

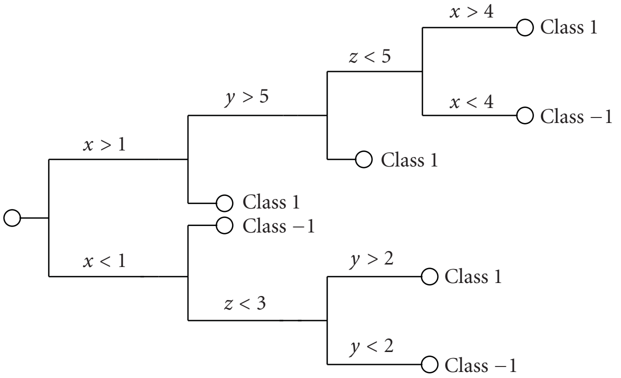


Figure 3: A sketch of a classification tree, wherein the leaves represent classes and the branches represent conjunctions of features that lead to those classes.

$$(\mathbf{x}, \mathbf{x}') \longmapsto k(\mathbf{x}, \mathbf{x}'). \tag{12}$$

The function $k(\cdot,\cdot)$ is called the *kernel function* and the so-called *kernel trick* uses Mercer's condition, which states that any positive semidefinite kernel $k(\mathbf{x}, \mathbf{x}')$ can be expressed as a dot product in a highdimensional space (see [31] for a detailed introduction). The theoretical foundations of this approach were given by Vapnik's *statistical learning theory* [6, 32] and later extended to the nonlinear case [7]. We use an implementation of SVMs that is based on the libsvm provided by Chang and Lin [33] with the standard kernels:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= (\mathbf{x} \cdot \mathbf{x}') \text{ linear} \\ &= (\mathbf{x} \cdot \mathbf{x}' + 1)^d \text{ polynomial} \\ &= \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{\sigma^2}\right) \text{ rbf} \end{aligned} \tag{13}$$

The parameters of the model are, with respect to the kernel-type, the polynomial degree $d$, the width of the rbf $\sigma^2$ and the value concerning the cost of constrain violation during the SVM training.

### 5.5. *Trees*

Trees are conceptually simple but powerful tools for classification and regression. For our purpose, we use the *classification and regression trees* (CARTs) as described in Breiman et al. [4]. The main feature of the CART algorithm is the binary decision role that is introduced at each tree node with respect to the information content of the split. In this way, the most discriminating binary splits are near the tree root building an hierarchical decision scheme. A sketch of a decision tree is shown in Figure 3. It is known that trees have a high variance, so they benefit from the ensemble approach [20]. These trees ensembles are also know as *random forests*. The parameters of the tree models are related to splitting the tree nodes (the impurity measure and the split criterion, see [2] for a detailed description).

### 5.6. Nearest-neighbor classifier

A $K$-nearest-neighbor classifier (KNN) takes a weighted average over the labels $z_i$ of those observations $\mathbf{z}_i$ in the training set that are closest to the query point $\mathbf{x}$. This denotes as

$$f(\mathbf{x}) = \frac{1}{\sum w_i} \sum_{\mathbf{z}_i \in N_k(\mathbf{x})} w_i z_i, \tag{14}$$

where $N_k(\mathbf{x})$ denotes the $k$-element neighborhood of $\mathbf{x}$, defined in a given metric, and $w_i$ is the related distance. Common choices are the $L_1$, $L_2$, and the $L_\infty$ metrics. The parameters of the model are the number of neighbors and the choice of the metric. KNNs offer a very intuitive approach to classification problems because they are based on the concept of similarity. This works fine in lower dimensions but leads to problems in higher dimensions, known as the *curse of dimensionality* [34].

## 6. APPLICATION TO THE CLINICAL DATA

We compared the model classes described above in a unified framework under fair conditions. Thus, we trained an ensemble of each model class consisting of 11 ensembles members (11 CV-folds in the training scheme described in Section 4). The performance of each ensemble model was assessed on the 20% of data (validation set), which was initially removed and never included in model training (see Figure 2). This procedure was independently repeated 20 times. This means that all model-building processes, that is, the random removal of 20% of the data, the construction of a classification model ensemble on the remaining 80% of the data as outlined in Section 4, and the final test on the unseen validation data were performed each time. Finally, the mean average prediction values with respect to the validation set were calculated and are listed in Table 2. In some cases, it is useful to apply a kind of data preprocessing like balancing. If the distribution of the two classes differ in the sense, that one class is only represented with a small number of examples, we can balance the data in the training set. This can improve the convergence of several training algorithms and has also an impact to the classification error [35]. We apply balancing in the way that we reduce the number of samples in the one class until we have an balanced ratio of the class labels. The ratio of the class labels in the validation set was never changed because it reflects the real data distribution. Balancing was only applied to the training data. We used three different performance measures in order to compare the different classification models. Therefore, we have to define the four possible outcomes of a classification that can be formulated in a 2 × 2 confusion matrix, as shown in Table 1. The accuracy,

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}, \tag{15}$$

seems to be the canonical error measure for almost all classification problems if the dataset is balanced. Other important measures are the specificity that quantifies how well a binary classification model correctly identifies the negative cases (non-PCa patients),

Table 1: The confusion matrix for a binary classification problem.

| | predicted class + 1 | predicted class − 1 |
|---|---|---|
| Real class + 1 | True positive (tp) | False negative (fn) |
| Real class − 1 | False positive (fp) | True negative (tn) |

$$\text{Specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}}, \tag{16}$$

and the sensitivity, which is the proportion of true positives of all diseased cases (PCa patients) in the population,

$$\text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{17}$$

A high sensitivity is required when early diagnosis and treatment are beneficial, which is the case in PCa.

The precision or positive predictive value (PPV) is given by

$$\text{PPV} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \tag{18}$$

and is the proportion of patients with positive test results who are correctly diagnosed. The F-Score is the harmonic mean of precision and sensitivity,

$$\text{F-Score} = 2 \cdot \frac{\text{Sensitivity} \cdot \text{PPV}}{\text{Sensitivity} + \text{PPV}}, \tag{19}$$

and it is useful if the classes in the classification problem are not equally distributed. Another measure is the area under curve (AUC) wherein the curve is the receiver operating characteristic (ROC-curve)curve. The ROC-curve is the graphical plot of the sensitivity versus the (1-specificity) for a binary classifier as its discrimination threshold is varied.

The ROC-curve offers the opportunity to calculate the specificity at a fixed sensitivity level and vice versa. This is important because, from the clinical point of view, a high sensitivity 95% is wanted to detect all patients with PCa first. To avoid a high false-positive rate, we computed the specificity at the level of 95% sensitivity (SPS95) from the ROC-curve as another important performance measure.

To have an impression about the correct classified non-PCa patients in this case, we computed the specificity at the level of 95% sensitivity (SPS95) from the ROC-curve. If we compare the outcome of the statistical analysis of the model performance as listed in Table 1 for the unbalanced case and in Table 2 for the balanced case, we can state that the differences between the different classifiers are marginal. Even the more sophisticated classification models (SVMs or Mixed Ensembles) could not outperform the robust linear candidates (LDA/PDA).

Tables 2 and 3 present the main results with only small differences between the classifiers. The standard deviations of the performance measures are given except for the ROC-curve-based measures (AUC and SPS95). Most papers in

Table 2: The average performance of several classifier ensembles with respect to the validation set which was initially removed and never included in model training. We show the mean and the standard deviation values from 20 independent validation runs, no preprocessing was used.

| | Accuracy | F-score | AUC | SPS95 |
|---|---|---|---|---|
| PDA | 0.776 ± 0.026 | 0.823 ± 0.026 | 0.863 | 0.454 |
| Log.Reg. | 0.778 ± 0.038 | 0.823 ± 0.036 | 0.868 | 0.484 |
| MLP | 0.791 ± 0.045 | 0.823 ± 0.04 | 0.863 | 0.453 |
| SVM | 0.795 ± 0.023 | 0.833 ± 0.02 | 0.825 | 0.142 |
| CART | 0.757 ± 0.03 | 0.809 ± 0.026 | 0.843 | 0.394 |
| KNN | 0.756 ± 0.036 | 0.813 ± 0.032 | 0.809 | 0.309 |
| Mixed | 0.783 ± 0.03 | 0.828 ± 0.026 | 0.860 | 0.457 |

Table 3: The average performance of several classifier ensembles with respect to the validation set which was initially removed and never included in model training. We show the mean and the standard deviation values from 20 independent validation runs wherein the training data was balanced.

| | Accuracy | F-score | AUC | SPS95 |
|---|---|---|---|---|
| PDA | 0.772 ± 0.034 | 0.809 ± 0.035 | 0.861 | 0.414 |
| Log.Reg. | 0.792 ± 0.03 | 0.834 ± 0.027 | 0.868 | 0.458 |
| MLP | 0.766 ± 0.027 | 0.787 ± 0.029 | 0.858 | 0.451 |
| SVM | 0.786 ± 0.038 | 0.816 ± 0.042 | 0.821 | 0.051 |
| CART | 0.755 ± 0.031 | 0.792 ± 0.029 | 0.841 | 0.376 |
| KNN | 0.726 ± 0.032 | 0.766 ± 0.034 | 0.801 | 0.297 |
| Mixed | 0.789 ± 0.033 | 0.830 ± 0.026 | 0.867 | 0.445 |

this field do not discuss this really complex problem and it cannot be solved in the scope of this paper, but it should be mentioned. As an example of a special solution of this problem, see the paper of Hilgers [36].

## 7. CONCLUSIONS

We compared several classification models with respect to their ability to recognize prostate cancer in an early stage. This was done in an ensemble framework in order to estimate proper model parameters and to increase classification performance. It turned out that all models under investigation are performing very well with only marginal differences and are compareable with similar studies, like, for example, Finne et al. [37], Remzi et al. [38], or Zlotta et al. [39]. In future research, it should be investigated whether these results are valid for other populations of patients (e.g., screening data) and other PSA test assays and whether the performance of classification could be increased by including new variables or by splitting the groups of patients into different PSA ranges.

## REFERENCES

[1] A. Jemal, R. Siegel, E. M. Ward, and M. J. Thun, “Cancer facts & figures,” Tech. Rep., Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Ga, USA, 2006.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, NY, USA, 2001.

[3] Y. Bard, *Nonlinear Parameter Estimation*, Academic Press, New York, NY, USA, 1974.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, Calif, USA, 1993.

[5] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1999.

[7] B. E. Boser, I. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the 5th Annual ACM Conference on Computational Learning Theory (COLT ’92)*, pp. 144–152, Pittsburgh, Pa, USA, July 1992.

[8] C. Merkwirth, U. Parlitz, and W. Lauterborn, “Fast nearest-neighbor searching for nonlinear signal processing,” *Physical Review E*, vol. 62, no. 2, pp. 2089–2097, 2000.

[9] J. D. Wichard and C. Merkwirth, “ENTOOL—A Matlab toolbox for ensemble modeling,” http://www.j-wichard.de/entool/, 2007.

[10] C. Stephan, H. Cammann, A. Semjonow, et al., “Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies,” *Clinical Chemistry*, vol. 48, no. 8, pp. 1279–1287, 2002.

[11] C. Stephan, M. Klaas, C. Müller, D. Schnorr, S. A. Loening, and K. Jung, “Interchangeability of measurements of total and free prostate-specific antigen in serum with 5 frequently used assay combinations: an update,” *Clinical Chemistry*, vol. 52, no. 1, pp. 59–64, 2006.

[12] M. P. Perrone and L. N. Cooper, “When networks disagree: ensemble methods for hybrid neural networks,” in *Neural Networks for Speech and Image Processing*, R. J. Mammone, Ed., pp. 126–142, Chapman-Hall, New York, NY, USA, 1993.

[13] A. Krogh and P. Sollich, "Statistical mechanics of ensemble learning," *Physical Review E*, vol. 55, no. 1, pp. 811–825, 1997.

[14] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7, pp. 231–238, MIT Press, Cambridge, Mass, USA, 1995.

[15] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[16] E. B. Kong and T. G. Dietterich, "Error-correcting output coding corrects bias and variance," in *Proceedings of the 12th International Conference on Machine Learning (ICML '95)*, pp. 313–321, Tahoe City, Calif, USA, July 1995.

[17] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proceedings of the 13th International Conference on Machine Learning (ICML '96)*, L. Saitta, Ed., pp. 275–283, Morgan Kaufmann, Bari, Italy, July 1996.

[18] P. Domingos, "A unified bias-variance decomposition for zero-one and squared loss," in *Proceedings of the 17th National Conference on Artificial Intelligence*, pp. 564–569, Austin, Tex, USA, July-August 2000.

[19] U. Naftaly, N. Intrator, and D. Horn, "Optimal ensemble averaging of neural networks," *Network: Computation in Neural Systems*, vol. 8, no. 3, pp. 283–296, 1997.

[20] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[21] C. Schaffer, "Selecting a classification method by cross-validation," in *Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics*, pp. 15–25, Fort Lauderdale, Fla, USA, January 1993.

[22] J. R. Quinlan, "Comparing connectionist and symbolic learning methods," in *Computational Learning Theory and Natural Learning Systems*, vol. 1, pp. 445–456, MIT Press, Cambridge, Mass, USA, 1994.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[24] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[25] J. D. Wichard, C. Merkwirth, and M. Ogorzałek, "Detecting correlation in stockmarkets," *Physica A*, vol. 344, no. 1-2, pp. 308–311, 2004.

[26] A. Rothfuss, T. Steger-Hartmann, N. Heinrich, and J. D. Wichard, "Computational prediction of the chromosome-damaging potential of chemicals," *Chemical Research in Toxicology*, vol. 19, no. 10, pp. 1313–1319, 2006.

[27] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, New York, NY, USA, 1989.

[28] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm ," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, pp. 586–591, San Francisco, Calif, USA, March-April 1993.

[29] C. Igel and M. Hüsken, "Improving the Rprop learning algorithm," in *Proceedings of the 2nd International ICSC Symposium on Neural Computation (NC '02)*, H. Bothe and R. Rojas, Eds., pp. 115–121, Academic Press, Berlin, Germany, May 2000.

[30] C. Merkwirth, J. D. Wichard, and M. Ogorzałek, "Stochastic gradient descent training of ensembles of DT-CNN classifiers for digit recognition," in *Proceedings of the 16th European Conference on Circuit Theory and Design (ECCTD '03)*, vol. 2, pp. 337–341, Kraków, Poland, September 2003.

[31] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[32] V. N. Vapnik and A. J. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie, Berlin, 1979.

[33] C. C. Chang and C. J. Lin, "Libsvm—Alibrary for support vector machines," 2001.

[34] R. E. Bellman, *Adaptive Control Processes*, Princeton University Press, Princeton, NJ, USA, 1961.

[35] C. Merkwirth, M. Ogorzałek, and J. D. Wichard, "Stochastic gradient descent training of ensembles of DT-CNN classifiers for digit recognition," in *Proceedings of the 16th European Conference on Circuit Theory and Design (ECCTD '03)*, vol. 2, pp. 337–341, Kraków, Poland, September 2003.

[36] R. A. Hilgers, "Distribution-free confidence bounds for ROC curves," *Methods of Information in Medicine*, vol. 30, no. 2, pp. 96–101, 1991.

[37] P. Finne, R. Finne, A. Auvinen, et al., "Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network," *Urology*, vol. 56, no. 3, pp. 418–422, 2000.

[38] M. Remzi, T. Anagnostou, V. Ravery, et al., "An artificial neural network to predict the outcome of repeat prostate biopsies," *Urology*, vol. 62, no. 3, pp. 456–460, 2003.

[39] A. R. Zlotta, M. Remzi, P. B. Snow, C. C. Schulman, M. Marberger, and B. Djavan, "An artificial neural network for prostate cancer staging when serum prostate specific antigen is 10 ng./ml. or less," *Journal of Urology*, vol. 169, no. 5, pp. 1724–1728, 2003.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 165730, 9 pages
doi:10.1155/2008/165730

*Research Article*

# Influence of Muscle-Tendon Wrapping on Calculations of Joint Reaction Forces in the Equine Distal Forelimb

**Jonathan S. Merritt,[1] Helen M. S. Davies,[2] Colin Burvill,[1] and Marcus G. Pandy[1]**

[1] *Department of Mechanical and Manufacturing Engineering, Melbourne School of Engineering, The University of Melbourne, VIC 3010, Australia*
[2] *Department of Veterinary Clinic and Hospital, Faculty of Veterinary Science, The University of Melbourne, VIC 3030, Australia*

Correspondence should be addressed to Jonathan S. Merritt, merritt@unimelb.edu.au

Received 14 September 2007; Accepted 24 December 2007

Recommended by Daniel Howard

The equine distal forelimb is a common location of injuries related to mechanical overload. In this study, a two-dimensional model of the musculoskeletal system of the region was developed and applied to kinematic and kinetic data from walking and trotting horses. The forces in major tendons and joint reaction forces were calculated. The components of the joint reaction forces caused by wrapping of tendons around sesamoid bones were found to be of similar magnitude to the reaction forces between the long bones at each joint. This finding highlighted the importance of taking into account muscle-tendon wrapping when evaluating joint loading in the equine distal forelimb.

## 1. INTRODUCTION

Mechanical loads experienced by the coffin (distal interphalangeal) and fetlock (metacarpophalangeal) joints of horses are thought to be related to injuries that occur in these joints. Mechanical stresses are thought to play a role in cartilage wear [1] and are likely to be the cause of many changes to subchondral bone that are associated with lameness [2]. Tendon wrapping is also thought to relate to the mechanics of injuries. For example, the influence of the deep digital flexor tendon (DDFT) on navicular disease [3, 4] and the relationship between interosseous ligament (IL) injury and metacarpal injury [5] are both presumed to be mediated by the forces generated when the two tendinous structures wrap around underlying bones. Consequently, knowledge of joint loads is likely to be relevant in understanding the mechanical pathogenesis of many kinds of injury in horses.

Little is currently known about the joint reaction forces of the equine distal forelimb. The wrapping of the DDFT about the distal sesamoid (navicular) bone has been investigated in very few studies [3, 4, 6–8]. Biewener et al. [9] calculated reaction forces at the coffin and fetlock joints, but those authors did not take into account forces produced by wrapping of the tendons around the sesamoid bones. Thomason [10] studied the architecture of the distal third metacarpal bone and estimated approximate values for the force between the proximal sesamoid bones and the distal metacarpus, but he did not perform any calculations based upon a dynamic model of the limb.

The aim of this study was to calculate and report the major joint reaction force components in the fetlock and coffin joints during walking and trotting in the horse. The forces exerted by wrapping of the tendons around both the proximal and distal sesamoid bones were taken into account when performing the joint reaction force calculations. The main hypothesis was that the forces exerted due to tendon wrapping would contribute substantially to the net reaction forces at the joints and hence could not be neglected.

## 2. MATERIALS AND METHODS

### 2.1. *Kinetic and kinematic data*

Three Quarter Horses (*Equus caballus*) with masses of 500 kg, 545 kg, and 500 kg were used for the study. Three retroreflective markers were attached to the skin over bony

Table 1: Parameters for the distal forelimb model. The $x$ axis is directed from the proximal joint of the given segment toward the distal joint and the $y$ axis is directed cranially. $x$ and $y$ values, defining either the origin and insertion points of muscle-tendon paths, or the centers of pulleys and via points, are given as percentages of their respective segment lengths.

| Tendon | Element | Segment | Radius (mm) | $x$ (%) | $y$ (%) |
|---|---|---|---|---|---|
| SDFT | Pulley | Metacarpus | 40.1 | 3.4 | 7.2 |
| | Via Point | Pastern | — | −7.1 | −27.6 |
| | Insertion | Pastern | — | 74.5 | −13.9 |
| IL | Origin | Metacarpus | — | 21.6 | 0.0 |
| | Pulley | Pastern | 19.3 | 0.0 | 0.0 |
| | Insertion | Pastern | — | 74.5 | −13.9 |
| DDFT | Origin | Metacarpus | 40.1 | 3.4 | 7.2 |
| | Via Point | Pastern | — | 1.6 | −31.0 |
| | Pulley | Hoof | 15.0 | 3.4 | −8.9 |
| | Insertion | Hoof | — | 32.0 | −11.6 |

parts of each segment of the right forelimb, where the segments were defined as the hoof, pastern, metacarpus, antebrachium, and brachium. Rather than being separated into portions corresponding to the first and second phalanges, the pastern was treated as a single rigid body because the proximal interphalangeal joint was previously described as approximately rigid [11]. Additional markers were then placed over the centers of the joints. A Peak Performance Motus system was used to measure the locations of the markers during a static trial, in which the horse stood quietly and approximately square.

The joint center markers were removed, and the horses were led in hand over a Bertec 4000 series force plate at the walk and trot. The ground reaction force was measured by the force plate, while kinematic data were simultaneously collected from the three markers of each limb segment. The horses were led so that one of the three orthogonal axes of the laboratory coordinate system corresponded to the direction of progression, and the other two axes were then presumed to be part of a parasagittal plane. A minimum of two trials and a maximum of four were collected for each horse at each gait, resulting in a total of 9 walking trials and 11 trotting trials. Marker position data were acquired at a frequency of 60 Hz, while ground reaction force data were acquired at 600 Hz.

### 2.2. Joint moment calculation

Kinematic and ground reaction force data from the walking and trotting trials were filtered using lowpass, forward-reverse Butterworth filters with mirror-symmetric boundary conditions. A cutoff frequency of 5 Hz was used for the kinematic data at the walk, while 12 Hz was used for the trot. The ground reaction force was filtered using a cutoff frequency of 12 Hz at both gaits. After filtering, the kinematic data were upsampled using cubic splines to a frequency of 600 Hz in order to match the sampling frequency of the ground reaction force. Noise was removed from the marker positions of the static trials by calculating the mean marker positions for each trial.

The joint center locations were calculated for the walking and trotting trials by referencing their locations during the static trial of each horse. The three markers of each limb segment that were common to both the static trial and the locomotion trials were used to calculate the joint center locations using three-dimensional rigid body transformations. For each sample of the locomotion trials, a transformation was calculated for each limb segment that moved the position of the joint center marker in the static trial to its expected position in the locomotion trial, under the assumption that the limb segment was a rigid body [12]. The center of each joint was calculated as the average of the positions predicted by the segment proximal to the joint and the segment distal to the joint. The two-dimensional, sagittal plane joint angles of the limb during each locomotion trial were then calculated after projecting the joint center locations to the sagittal plane.

The sagittal plane joint moments [13] were calculated for the stance phase of each locomotion trial, defined as the portion of the trial during which the vertical ground reaction force was greater than 50 N. The joint moments were calculated using two different methods: standard inverse dynamics and a massless, quasistatic analysis. In the first method, the inertial parameters (mass and mass moment of inertia) of each limb segment were approximated using regression equations determined for a set of Dutch Warmblood horses [14]. For the second method, both the gravitational mass and inertial parameters of each segment were set equal to zero.

### 2.3. Forelimb model

A two-dimensional model of the distal forelimb was developed, based upon the model described by Meershoek et al. [15]. Brown et al. [16] performed tendon excursion measurements to find the moment arms of several muscles in the equine distal forelimb about the carpus and fetlock joints. The moment arm values measured by Brown et al. [16] were substantially different from those of the model described by Meershoek et al. [15], and because of this discrepancy, the model was modified so that its moment arms matched those reported by Brown et al. [16] Figure 1 illustrates the modified forelimb model, while Table 1 describes the complete model geometry. A comparison of the moment arms of the superficial digital flexor tendon(SDFT) and DDFT at the fetlock joint, as reported by both Meershoek et al. [15] and Brown et al. [16], yielded a scaling factor of 0.68. All pulley radii in the Meershoek et al. [15] model were scaled by this value.

Uniform scaling of pulley radii did not yield a good match at the fetlock for the variation in moment arm with joint angle as reported by Brown et al. [16]. In order to more correctly simulate this variation, the pulleys of the SDFT and DDFT were replaced by a single via point for each tendon, whose location represented a point of wrapping around the proximal sesamoid bones. The paired proximal sesamoid bones are connected to the first phalanx by their straight, oblique, and short distal ligaments [17]. These relatively short, strong ligaments are unlikely to experience considerable length changes during locomotion [10], and hence the wrapping locations of the tendons around the fetlock joint were approximated as fixed relative to the pastern. The locations of the via points were chosen to correspond to the approximate anatomical locations of the sesamoid bones, and were then fine-tuned to produce moment arms which closely matched those reported by Brown et al. [16]. The comparison between model moment arms and the values published by Brown et al. [16] is shown in Figure 2.

### 2.4. Tendon tension and joint reaction force calculations

The forelimb model was used to calculate the tensions in the SDFT, DDFT, and IL, as well as the reaction forces at the coffin and fetlock joints. Following the method described by Meershoek et al. [15], the strain in the IL was first calculated assuming that the rest length of this ligament was its length at the start of the stance phase [15]. The tension in the IL was calculated using an experimentally-derived force-length relationship [15]. The tension of the DDFT was then calculated by balancing the joint moment at the coffin joint with the moment generated by this tendon. A similar calculation was performed at the fetlock joint, to calculate the tension in the SDFT, after subtracting the moments exerted about this joint by both the IL and the DDFT.

Finally, joint reaction forces were calculated at the coffin and fetlock joints by enforcing static equilibrium of all forces acting on the segments shown in Figure 3. The DDFT wrapped around the distal sesamoid (navicular) bone, causing it to be compressed against the distal palmar articular surface of the second phalanx [4]. In a similar way, the SDFT, DDFT, and IL wrapped around the proximal sesamoid bones, causing them to be compressed against the palmar surface of the distal metacarpal condyles. These wrapping forces were considered as part of the calculation of joint reaction forces.

The duration of the stance phase of each trial varied slightly. Consequently, after all quantities had been calculated, the stance phase duration was normalized to a percentage scale, with 0% representing the first hoof-ground contact and 100% representing the last hoof-ground contact. The time sample of each trial was scaled linearly to fit this range. Following this, the results were scaled by body mass, by dividing them by the mass of each horse expressed in kilograms. The results were then averaged over all trials to obtain estimates of the mean and standard deviation of each quantity. Due to the small number of trials available, all trials

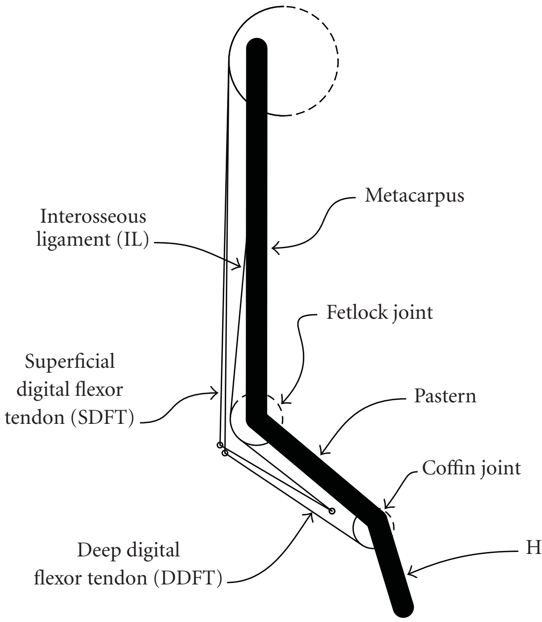

Metacarpus
Interosseous ligament (IL)
Fetlock joint
Superficial digital flexor tendon (SDFT)
Pastern
Coffin joint
Deep digital flexor tendon (DDFT)
Hoof

Figure 1: Two-dimensional, sagittal plane model of the equine forelimb. The model contained three rigid body segments and three major tendinous structures.

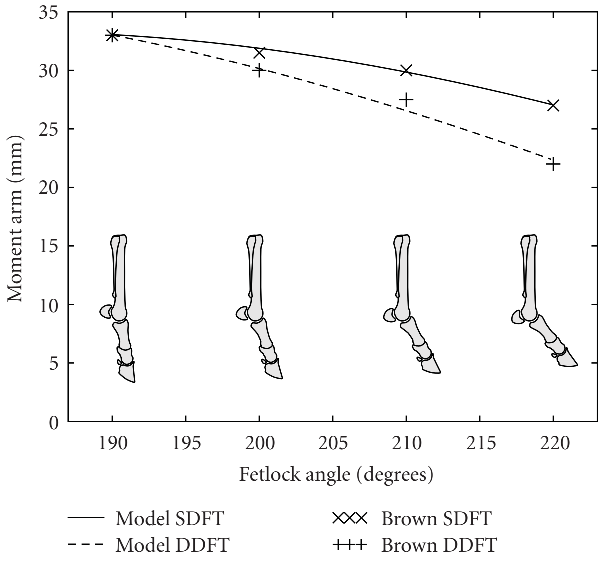


Figure 2: Moment arms from the forelimb model compared with those published by Brown et al. [16]. Moment arms for the forelimb model were calculated using a simulated tendon excursion method.

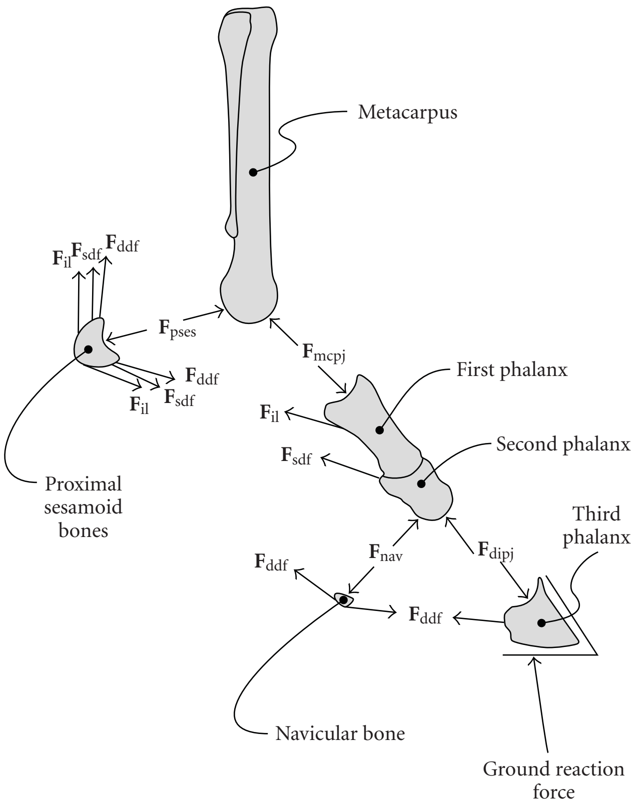


Figure 3: Forces considered when calculating the joint reaction forces at the fetlock and coffin joints. The forces were the ground reaction force, the force exerted by the distal phalanx on the middle phalanx ($\mathbf{F}_{dipj}$), the force exerted by the navicular bone on the middle phalanx ($\mathbf{F}_{nav}$), the DDFT force ($\mathbf{F}_{ddf}$), the SDFT force ($\mathbf{F}_{sdf}$), the IL force ($\mathbf{F}_{il}$), the force exerted by the proximal phalanx on the distal metacarpus ($\mathbf{F}_{mcpj}$), and the force exerted by the proximal sesamoid bones on the distal metacarpus ($\mathbf{F}_{pses}$).

were weighted equally in this calculation, despite the fact that some horses performed more trials than others.

## 3. RESULTS AND DISCUSSION

The model presented in this paper had several important limitations. The two-dimensional nature of the model limited it to considering only those force components which could be projected onto the sagittal plane. This approximation may affect all of the calculated quantities, especially where forces are oriented so as to be partially oblique to a parasagittal plane. Additionally, frontal asymmetry of the limb, both in its loading and anatomy, may influence the internal loads significantly. The importance of frontal plane asymmetry is suggested by many clinical phenomena; for example, the greater reported occurrence of proximal sesamoid bone fracture in the medial bone compared with the lateral one in racing Thoroughbreds [18]. The model presented in this paper presumed perfect frontal symmetry, and hence would be unable to predict the effects of different loading or anatomy of the medial and lateral sides of the limb.

Only the major anatomical structures of the distal limb were modeled. In reality, the proximal sesamoid bones are not fixed relative to the pastern, and additional structures such as the distal sesamoidean ligaments, the collateral ligaments of the joints, and the extensor branches of the suspensory (interosseous) ligament are all likely to exert loads on the bones, which would change the estimates of joint loading obtained in this study. The decision to include only the major structures of the limb was motivated both by a desire for simplicity in the model and a lack of data regarding the morphology and mechanical properties of the remaining structures. The justification for this decision lies mainly in precedent, for example, [15], and the anatomical observation that the structures that were included in the model were physically much larger than those which were neglected.

The exact zero strain (rest) length of IL of the model was unknown, and so it was approximated as the length of the IL during early stance, following the method described by Meershoek and Lanovaz [19]. It was observed that in the model, the IL behaved as a nonlinear angular spring, where the moment it exerted about the fetlock was dependent only upon the fetlock angle. Hence, variations in the rest length of IL would affect the relative sharing of the moment at the fetlock joint between the IL and SDFT. It would be beneficial for a future study to investigate directly the relationship between the IL moment about the fetlock and the fetlock joint angle, rather than inferring the moment about this joint from angle-strain and strain-force relationships.

The model simulated the moment arms of the SDFT and DDFT at the fetlock that were reported by Brown et al. [16]. These moment arms were measured by the tendon excursion method, and were substantially different from those reported previously in the literature [9, 15, 20]. The reason for the discrepancy in moment arm values is not known, but may arise from different measurement methodologies or different breeds of horse being used. Further investigation would be beneficial to the development of mechanical models of the equine forelimb in the future. Variation in the SDFT and DDFT moment arms at the fetlock could affect the calculated tension in the SDFT.

Figure 4 shows the joint moments calculated at the coffin, fetlock, carpus, and elbow joints for walking and trotting. The calculated joint moments were comparable with those reported previously in the literature [21, 22]. These two previous studies [21, 22] used Dutch Warmblood horses, whereas the current study used Quarter Horses. The difference in horse breeds may account for small differences in the joint moments.

In Figure 4, it is evident that the massless, quasistatic calculation of joint moments produced results very similar to the full inverse dynamics calculation for all joints. To the best knowledge of the authors, the direct comparison of these two techniques for calculating joint moments in the equine distal forelimb has never been reported, despite the reliance on quasistatic solutions for several previous distal forelimb studies, for example, [23]. The relative importance of the inertial and mass-related components of the moments increased for the more proximal joints of the limb, where greater masses and greater accelerations were involved. The

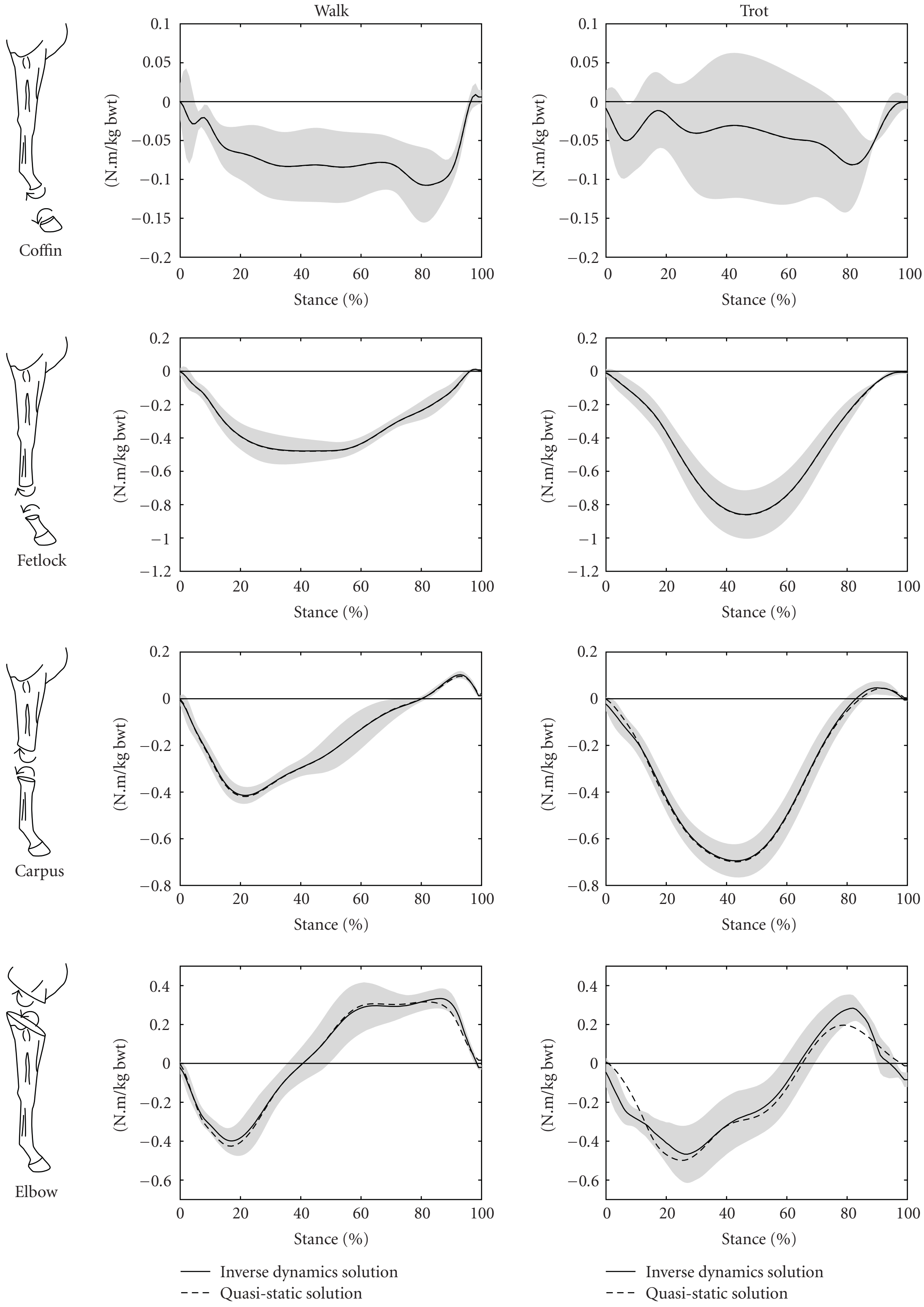


Figure 4: Joint moments calculated at the coffin, fetlock, carpus, and elbow joints for walking and trotting during the stance phase of the stride. The mean values of the inverse dynamics solutions are shown as solid lines with grey shading representing ±1 SD, while the mean values of the massless, quasi-static analysis are shown as dashed lines. The stance phase of the stride was the time during which the vertical component of the ground reaction force was greater than 50 N. 0% of stride was the time of first hoof-ground contact, while 100% of stride was the time of last hoof-ground contact.

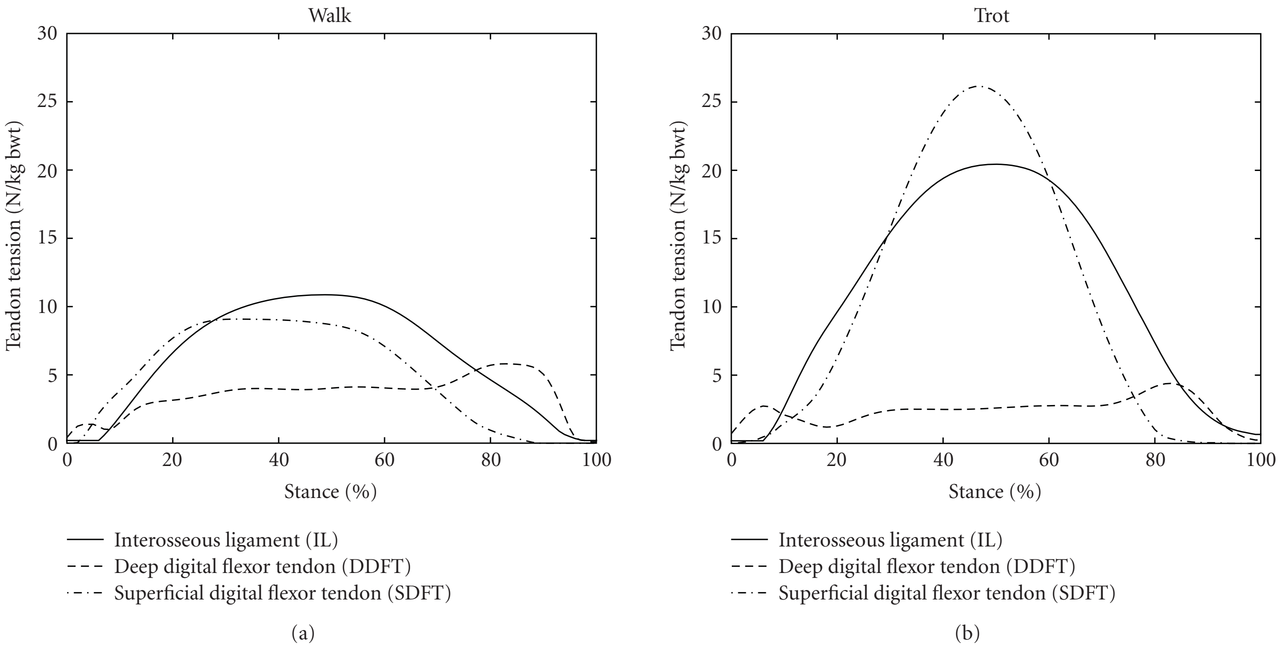


Figure 5: Mean tendon tensions in the DDFT, SDFT, and IL for walking and trotting during the stance phase of the stride. The stance phase of the stride was the time during which the vertical component of the ground reaction force was greater than 50 N. 0% of stride was the time of first hoof-ground contact, while 100% of stride was the time of last hoof-ground contact.

insensitivity of the results to inertial factors during the stance phase contrasted with the reported importance of these factors during the swing phase [24]. This contrast may readily be understood as a result of the ground reaction force, which dominated the joint moment calculations during the stance phase, yet is absent by definition during the swing phase. In the present model, each forelimb distal to the carpus accounted for only 0.6% of body mass [14], and this region of the limb experienced little acceleration during stance. Hence it was unlikely that any inertial or mass-related effects from this region of the limb would have a significant effect on calculated joint moments or other aspects of limb dynamics.

The mean tensions in the DDFT, SDFT, and IL are shown in Figure 5 for walking and trotting. The tendon tensions during walking were similar to those calculated by Jansen et al. [25], from in vivo tendon strain measurements. The tensions during trotting were somewhat different from those previously reported by Meershoek and Lanovaz [19], but the differences may be accounted for by the substantially different moment arms of the tendons at the fetlock in the two studies. In the present study, the tendon moment arms at the fetlock were configured to simulate values determined experimentally using the tendon excursion method [16], while those used by Meershoek and Lanovaz [19] were obtained using a geometric method [15]. The rest length of IL was also unknown in the present study, and so it was approximated using the method described by Meershoek and Lanovaz [19]. Variations in the IL rest length would cause changes in the IL and SDFT tensions.

Magnitudes of the components of the joint reaction forces are shown in Figure 6 for walking and trotting. Shown are the force exerted by the distal phalanx on the middle phalanx, the force exerted by the navicular bone on the middle phalanx, the force exerted by the proximal phalanx on the distal metacarpus, and the force exerted by the paired proximal sesamoid bones on the distal metacarpus.

In a previous report of joint reaction forces in the equine distal forelimb, Biewener et al. [9] noted that forces may be generated from tendon wrapping, but these authors did not calculate those forces and ignored them in their subsequent analysis of bone strains. Thomason [10] estimated the magnitude of the wrapping force at the proximal sesamoid bones from assumptions derived from studies of the architecture of the distal third metacarpal bone, but he did not perform any calculations based upon a dynamic model of the limb. Several previous studies [3, 4, 6–8] calculated the wrapping force of the DDFT about the distal sesamoid bone, and a link has been postulated between this force and navicular disease. However, to the best knowledge of the authors, the remaining joint reaction results represent novel data which have not previously been reported.

The calculated force exerted by the navicular bone on the middle phalanx was similar to that reported previously [3, 4, 6–8]. In particular, this force exhibited a peak in late stance, which was related to the change in the point of application (point of zero moment) of the ground reaction force [4]. It was observed that as the hoof neared breakover (the point in the stride at which the heels of the hoof depart from the ground and the hoof "break over" the toe), the point of application of the ground reaction force moved from a location within the body of the hoof to a point at the toe, which increased the moment arm of this force about the coffin joint. During late stance, this increase in moment arm

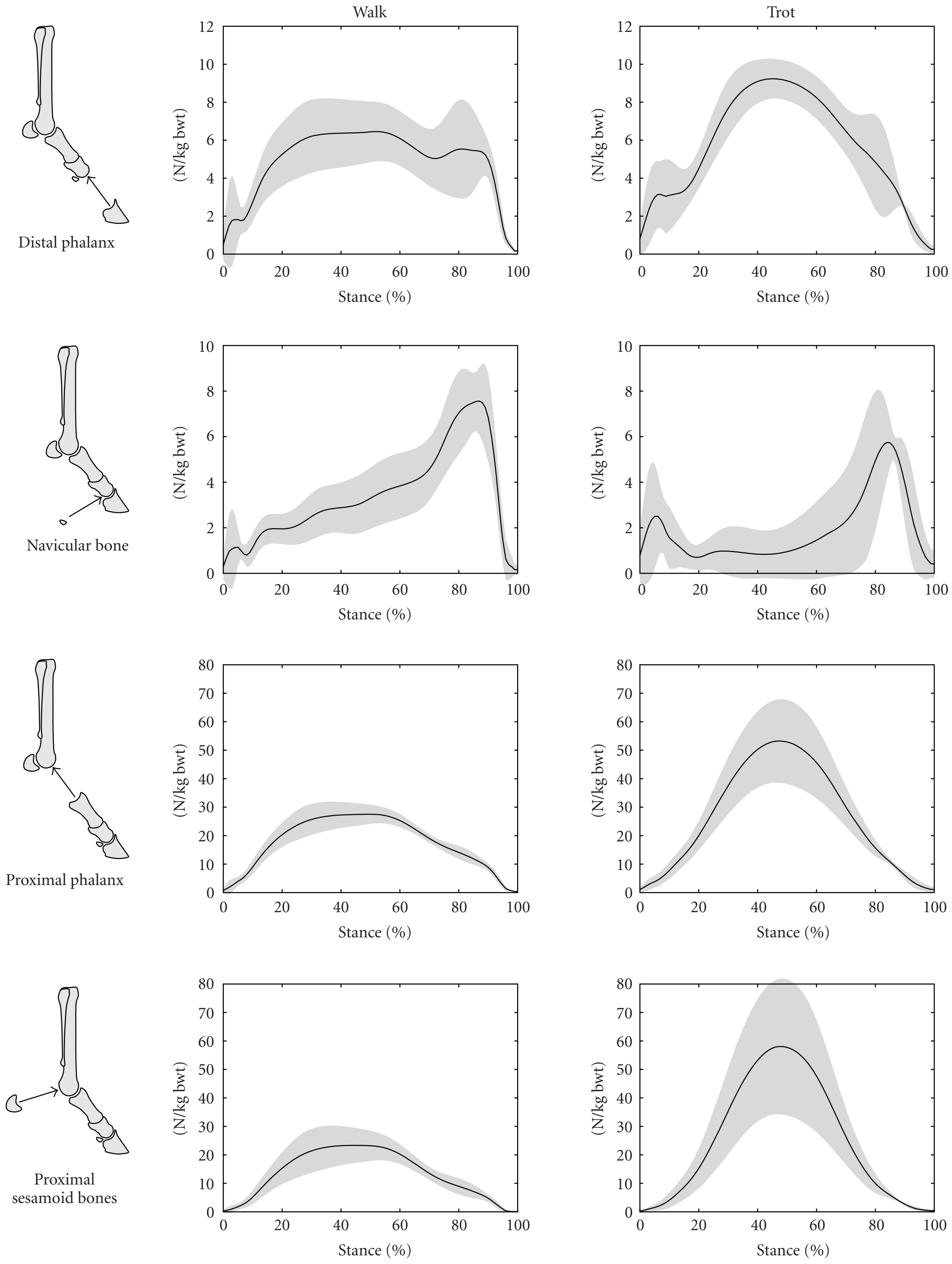


Figure 6: Magnitudes of calculated joint reaction force components from long bones and sesamoid bones in the coffin and fetlock joints for walking and trotting during the stance phase of the stride. The magnitudes are shown as solid lines with grey shading to represent ±1 SD. The stance phase of the stride was the time during which the vertical component of the ground reaction force was greater than 50 N. 0% of stride was the time of first hoof-ground contact, while 100% of stride was the time of last hoof-ground contact.

was adequate to raise the value of the coffin joint moment (Figure 4, Coffin), DDFT tension (Figure 5, DDFT), and the force applied by the navicular bone (Figure 6, Navicular Bone) despite a decrease in the magnitude of the ground reaction force in late stance relative to midstance (observed in the ground reaction force data, but not shown here). The increase in DDFT tension in late stance did not produce such an obvious peak in the reaction force between the distal and middle phalanges (Figure 6, Distal Phalanx) although a bimodal pattern was evident in this component during walking (Figure 6, Distal Phalanx, Walk).

The central hypothesis of this paper—that the forces induced by wrapping of tendons around the sesamoid bones are important components of the net joint reaction forces of the distal forelimb—is well illustrated in Figure 6. In the top two rows of Figure 6, the magnitude of the distal phalanx force and that of the navicular bone are both comparable, as are the magnitudes of the proximal phalanx and proximal sesamoid bone forces shown in the bottom two rows. Hence, the magnitudes of forces exerted by the sesamoid bones of the joints were comparable to those exerted between the long bones themselves. Neglect of either of the two forces from the sesamoid bones would cause a substantial change in the calculated net joint reaction force.

Knowledge of the reaction forces at the joints is potentially important for future biomechanical studies into areas such as the behavior of joint cartilage, subchondral bone loading, and bone remodeling. Future studies are expected to involve three-dimensional finite-element modeling of the mechanical interactions at the joints. The model presented in this paper, and its calculated joint reaction forces, could be used as a reference point for the predictions of more sophisticated models.

## 4. CONCLUSIONS

Joint reaction forces in the equine distal forelimb were calculated using a two-dimensional mathematical model. Components of these forces that were generated by wrapping of tendons about the sesamoid bones in the limb were found to be of similar magnitude to those generated between the long bones of each limb segment. These wrapping forces serve to illustrate the importance of the sesamoid bones in the mechanical function of the forelimb, and also provide insight into the possible mechanical cause of numerous injuries related to loading of the joints.

## ACKNOWLEDGMENTS

The authors would like to thank Drs Nick Brown, Chris Kawcak, and C. Wayne McIllwraith for the use of the kinematic and kinetic data from three Quarter Horses. These data were collected using facilities at Colorado State University.

## REFERENCES

[1] R. M. Bowker, P. J. Atkinson, T. S. Atkinson, and R. C. Haut, "Effect of contact stress in bones of the distal interphalangeal joint on microscopic changes in articular cartilage and ligaments," *American Journal of Veterinary Research*, vol. 62, no. 3, pp. 414–424, 2001.

[2] C. E. Kawcak, C. W. McIlwraith, R. W. Norrdin, R. D. Park, and P. S. Steyn, "Clinical effects of exercise on subchondral bone of carpal and metacarpophalangeal joints in horses," *American Journal of Veterinary Research*, vol. 61, no. 10, pp. 1252–1258, 2000.

[3] M. P. McGuigan and A. M. Wilson, "The effect of bilateral palmar digital nerve analgesia on the compressive force experienced by the navicular bone in horses with navicular disease," *Equine Veterinary Journal*, vol. 33, no. 2, pp. 166–171, 2001.

[4] A. M. Wilson, M. P. McGuigan, L. Fouracre, and L. MacMahon, "The force and contact stress on the navicular bone during trot locomotion in sound horses and horses with navicular disease," *Equine Veterinary Journal*, vol. 33, no. 2, pp. 159–165, 2001.

[5] S. S. Le Jeune, M. H. Macdonald, S. M. Stover, K. T. Taylor, and M. Gerdes, "Biomechanical investigation of the association between suspensory ligament injury and lateral condylar fracture in thoroughbred racehorses," *Veterinary Surgery*, vol. 32, no. 6, pp. 585–597, 2003.

[6] E. Eliashar, M. P. McGuigan, K. A. Rogers, and A. M. Wilson, "A comparison of three horseshoeing styles on the kinetics of breakover in sound horses," *Equine Veterinary Journal*, vol. 34, no. 2, pp. 184–190, 2002.

[7] E. Eliashar, M. P. McGuigan, and A. M. Wilson, "Relationship of foot conformation and force applied to the navicular bone of sound horses at the trot," *Equine Veterinary Journal*, vol. 36, no. 5, pp. 431–435, 2004.

[8] M. A. Willemen, H. H. C. M. Savelberg, and A. Barneveld, "The effect of orthopaedic shoeing on the force exerted by the deep digital flexor tendon on the navicular bone in horses," *Equine Veterinary Journal*, vol. 31, no. 1, pp. 25–30, 1999.

[9] A. A. Biewener, J. Thomason, A. Goodship, and L. E. Lanyon, "Bone stress in the horse forelimb during locomotion at different gaits: a comparison of two experimental methods," *Journal of Biomechanics*, vol. 16, no. 8, pp. 565–576, 1983.

[10] J. J. Thomason, "The relationship of structure to mechanical function in the third metacarpal bone of the horse, *Equus caballus*," *Canadian Journal of Zoology*, vol. 63, no. 6, pp. 1420–1428, 1985.

[11] C. Degueurce, H. Chateau, H. Jerbi, et al., "Three-dimensional kinematics of the proximal interphalangeal joint: effects of raising the heels or the toe," *Equine Veterinary Journal. Supplement*, no. 33, pp. 79–83, 2001.

[12] J. H. Challis, "A procedure for determining rigid body transformation parameters," *Journal of Biomechanics*, vol. 28, no. 6, pp. 733–737, 1995.

[13] W. Back and H. Clayton, Eds., *Equine Locomotion*, W.B. Saunders, Philadelphia, Pa, USA, 2000.

[14] H. H. F. Buchner, H. H. C. M. Savelberg, H. C. Schamhardt, and A. Barneveld, "Inertial properties of Dutch Warmblood horses," *Journal of Biomechanics*, vol. 30, no. 6, pp. 653–658, 1997.

[15] L. S. Meershoek, A. J. van den Bogert, and H. C. Schamhardt, "Model formulation and determination of in vitro parameters of a noninvasive method to calculate flexor tendon forces in the equine forelimb," *American Journal of Veterinary Research*, vol. 62, no. 10, pp. 1585–1593, 2001.

[16] N. A. T. Brown, M. G. Pandy, W. L. Buford, C. E. Kawcak, and C. W. McIlwraith, "Moment arms about the carpal and metacarpophalangeal joints for flexor and extensor muscles

in equine forelimbs," *American Journal of Veterinary Research*, vol. 64, no. 3, pp. 351–357, 2003.

[17] K. M. Dyce, W. O. Sack, and C. J. G. Wensing, *Textbook of Veterinary Anatomy*, W.B. Saunders, Philadelphia, Pa, USA, 1987.

[18] E. J. Parente, D. W. Richardson, and P. Spencer, "Basal sesamoidean fractures in horses: 57 cases (1980–1991)," *Journal of the American Veterinary Medical Association*, vol. 202, no. 8, pp. 1293–1297, 1993.

[19] L. S. Meershoek and J. L. Lanovaz, "Sensitivity analysis and application to trotting of a noninvasive method to calculate flexor tendon forces in the equine forelimb," *American Journal of Veterinary Research*, vol. 62, no. 10, pp. 1594–1598, 2001.

[20] A. J. van den Bogert, "Computer simulation of locomotion in the horse," Ph.D. thesis, The University of Utrecht, Utrecht, The Netherlands, 1989.

[21] H. M. Clayton, J. L. Lanovaz, H. C. Schamhardt, M. A. Willemen, and G. R. Colborne, "Net joint moments and powers in the equine forelimb during the stance phase of the trot," *Equine Veterinary Journal*, vol. 30, no. 5, pp. 384–389, 1998.

[22] G. R. Colborne, J. L. Lanovaz, E. J. Sprigings, H. C. Schamhardt, and H. M. Clayton, "Forelimb joint moments and power during the walking stance phase of horses," *American Journal of Veterinary Research*, vol. 59, no. 5, pp. 609–614, 1998.

[23] H. F. Schryver, D. L. Bartel, N. Langrana, and J. E. Lowe, "Locomotion in the horse: kinematics and external and internal forces in the normal equine digit in the walk and trot," *American Journal of Veterinary Research*, vol. 39, no. 11, pp. 1728–1733, 1978.

[24] J. L. Lanovaz and H. M. Clayton, "Sensitivity of forelimb swing phase inverse dynamics to inertial parameter errors," *Equine Veterinary Journal. Supplement*, no. 33, pp. 27–31, 2001.

[25] M. O. Jansen, A. J. van den Bogert, D. J. Riemersma, and H. C. Schamhardt, "In vivo tendon forces in the forelimb of ponies at the walk, validated by ground reaction force measurements," *Acta Anatomica*, vol. 146, no. 2-3, pp. 162–167, 1993.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 623475, 7 pages
doi:10.1155/2008/623475

*Research Article*

# Alternative Parametric Boundary Reconstruction Method for Biomedical Imaging

**Joseph Kolibal**[1] **and Daniel Howard**[2]

[1] *Department of Mathematics, College of Science and Technology, The University of Southern Mississippi, Hattiesburg, MS 39406-0001, USA*
[2] *QinetiQ PLC, Malvern, Worcestershire WR14 3PS, UK*

Correspondence should be addressed to Joseph Kolibal, joseph.kolibal@usm.edu

Received 2 September 2007; Accepted 15 January 2008

Recommended by Halima Bensmail

Determining the outline or boundary contour of a two-dimensional object, or the surface of a three-dimensional object poses difficulties particularly when there is substantial measurement noise or uncertainty. By adapting the mathematical approach of stochastic function recovery to this task, it is possible to obtain usable estimates for these boundaries, even in the presence of large amounts of noise. The technique is applied to parametric boundary data and has potential applications in biomedical imaging. It should be considered as one of several techniques to improve the visualization of images.

## 1. INTRODUCTION

Three-dimensional (3D) computer reconstruction of a target volume or of a surface is an important activity in modern biomedical imaging. The accurate anatomical reconstruction in trauma or for use in image-guided intervention relies on mathematical imaging technology; and this paper develops the mathematical technique of stochastic function recovery [1] and illustrates its use for noisy boundary reconstruction. This is an alternative approach to the standard polynomial-based methods that we see as an add-on or complement to other techniques in use or being developed to improve upon the reconstruction of noisy boundary data to provide enhanced biomedical visualization.

The ability to distinguish features related to boundaries is intrinsic to technology of visualization. For example, in MRI imaging, a range of specialized methods have been developed for extracting information from signals so as to reconstruct images representing internal body structures [2]. Boundary recovery techniques apply to complex surgical procedures as with electroanatomical mapping that tracks the position of catheters inside the body with sparse signals recorded from electrodes at the tip of the catheter. Resulting surface maps must integrate real-time measurements with preoperative MR or CT images, and account for mapping data errors in registration and error due to patient movement [3]. In general, when signals are affected by noise, it must be effectively removed in order to improve the visualization and compared with other medical imaging modalities, ultrasound images suffer from speckle noise that often makes for weak or incomplete boundaries [4].

Our approach is to use stochastic convolution-deconvolution operators [1, 5–7], which have useful statistical properties, to smooth noisy surface data in a manner that does not obliterate detail, and which effectively removes Gaussian noise. The motivation for the approach is that, intrinsically, stochastic interpolation using probabilistic kernels for the generating function of the row space of the linear operator performs well at removing noise when used to approximate data. However, the difficulty in applying these methods directly is that they bias the data to a mean of zero.

In approximating one-dimensional data, this is not usually a difficulty. However, in approximating multidimensional data, this can cause an apparent shift in the approximant when working in parametric coordinates. This is because the Gaussian kernels used smooth positive values to their mean value, thus potentially shifting the coordinates nearer the origin. When we approximate in one dimension, the data values are shifted to the mean, however, the coordinates are not touched. If we are using parametric data, the coordinates are the data values ; and this means that the data can be shifted in $(x, y)$ or $(x, y, z)$ space. The less data existing or the greater the smoothing, the more will be this shift.

For example, when data is taken from a circle centered on the origin, the approximating curve is a circular curve centered on the origin, but of smaller radius, and the greater the smoothing applied in the approximation (if the data is very noisy), the smaller the area circumscribed by the approximating curve. As the number of sample points is increased, the approximation improves. However, for coarse surface data that is only smoothly varying, this can cause difficulties.

One approach is to make use of stochastic interpolation. Creating a dense noisy data set from the sparse noisy data using interpolation is followed by approximating this fine data set to recover the smoothed surface curve, thereby mitigating the shifting of the mean. While workable, this approach has several disadvantages, most notably that it requires a more costly interpolation step. It also requires the application of the technique more than once, and in the second or subsequent applications, the approximation must be done on a fine data set, meaning that many more points require approximation, again incurring a larger computational cost.

The solution is to make use of the convolution-deconvolution properties of stochastic interpolation combining the densification step with the approximation step. This would still be expensive. However, we introduce an approximate means for doing the interpolation that interpolates for smooth data, but which approximates for noisy data, thus avoiding the costly need to construct the inverse operator needed for interpolation.

## 2. DEVELOPMENT

Consider the task of sampling a known function $f(u)$ at points $f(u_k) = v_k$ with $u_k \in [0, 1]$ so as to determine its value at $x_j \in [0, 1]$. The *stochastic interpolant* [5] to the data $\{(u_k, v_k)\}$, $k = 1, \dots, n$ is given by

$$B_{mn}A_{nn}^{-1}v, \tag{1}$$

where $v = (v_1, v_2, \dots, v_n)^T$ is the data vector, and where $A_{nn}$ is a row stochastic matrix whose coefficients consist of the $n \times n$ values $a_{jk}$. Choosing the generator of the row space of $A_{nn}$ to be the Bernstein functions [1] (named after Bernstein as the derivation of this form that can be obtained from the Bernstein polynomials), we have

$$a_{jk} = \frac{1}{2}\left[\operatorname{erf}\left(\frac{u_{k+1} - x_j}{\sqrt{\sigma n}}\right) + \operatorname{erf}\left(\frac{x_j - u_k}{\sqrt{\sigma n}}\right)\right] \tag{2}$$

with $\sigma > 0$ on the partition $w_k = (u_k + u_{k+1})/2$, with $w_0 = -\infty$ and $w_n = \infty$ yielding a stochastic matrix. Setting $x_j = u_j$ generates the entries of $A_{nn}$, and setting $x_j$ to any set of $m$ consecutive values in $[0, 1]$ generates the coefficients of $B_{mn}$, that is, the coefficients of $B_{mn}$ are constructed in the same manner as for $A_{nn}$, except that the nodes $x_j$ at which $b_{jk}$ is evaluated may differ from the values at which the data are given. While any probability density function (pdf) can be used, appropriately replacing the mean and variance of the Gaussian in (2), a pdf based on the normal distribution is consistent with the problem of filtering Gaussian noise.

In stochastic interpolation, with the coefficient of $A_{nn}$ generated by (2), we can interpret $A_{nn}^{-1}$ as the discrete deconvolution of the data yielding the preimage generated by $A_{nn}^{-1}v$. This preimage is then convolved by $B_{mn}$ to yield an $m$-vector of values that interpolates the data $v$ at the output coordinates $x_j$, $j = 1, \dots, m$; the output coordinates are those that were used to generate the coefficients of $B_{mn}$. It is for this reason that we have elected to represent the matrix $A$ using another symbol $B$ since it is desirable to emphasize its role in convolution.

Defining $w_k = (u_k + u_{k+1})/2$ with $w_0 = -u_0/2$ and $w_n = 1 + u_n/2$ with $\sigma$ constant yields a coefficient structure in which $A_{nn}$ is a diagonal matrix times a symmetric Toeplitz matrix: $A_{nn} = DT_{nn}$. Inversion or solution of these matrices can be accomplished in $\mathcal{O}(n^2)$ operations [8], however in the cases of interest to us, this is not necessary. It is possible to do better than this using an approximate inverse in which the row space of $A_{nn}^I \approx A_{nn}^{-1}$ is generated directly. While evaluating $B_{mn}A_{nn}^I$ is still $\mathcal{O}(n^2)$, it is significantly faster than the Toeplitz matrix inversion of $A_{nn}$ to obtain $A_{nn}^{-1}$.

The approximate inverse is an approximation precisely because $A_{nn}A_{nn}^I \neq I_{nn}$, that is, in applying stochastic interpolation to the data $v$, there is an error:

$$e_0 = v - A_{nn}A_{nn}^I v = (I_{nn} - A_{nn}A_{nn}^I)v. \tag{3}$$

Thus the interpolant to the data can be expressed using successive correction to the errors using

$$\begin{aligned} v &= A_{nn}A_{nn}^I v + e_0 \\ &= A_{nn}A_{nn}^I v + AA_{nn}^I e_0 + e_1 \\ &\ \ \vdots \\ &= A_{nn}A_{nn}^I v + A_{nn}A_{nn}^I e_0 + A_{nn}A_{nn}^I e_1 \\ &\quad + \cdots + A_{nn}A_{nn}^I e_p + e_{p+1}. \end{aligned} \tag{4}$$

Substituting for $e_k$, from $k = 0$ to $p$, gives

$$v = A_{nn}A_{nn}^I \left(\sum_{k=0}^{p} (I_{nn} - A_{nn}A_{nn}^I)^k\right) v + e_{p+1}, \tag{5}$$

and truncating the sum gives a working formula in which $\widetilde{v} \approx v$ even for larger values of $\sigma$, so that the method nearly interpolates. Truncating at $p = 2$ and applying the formula to generate $m$ output values instead of $n$ gives

$$\widetilde{v} = B_{mn}A_{nn}^I(I_{nn} + G_{nn} + G_{nn}^2)v, \tag{6}$$

where $G_{nn} = (I_{nn} - A_{nn}A_{nn}^I)$. Provided that $\sigma$ in (2) is small, the error in constructing $A_{nn}^{-1}$ using $A_{nn}^I$ is small, however if larger values of $\sigma$ are used, then greater smoothing is applied to the data and the use of (6) becomes necessary when $\sigma$ is larger than 0.05. However, it will become apparent that it is because of this greater smoothing that it is unnecessary to apply any corrections as shown in (6), and thus the direct computation of $B_{mn}A^I_{nn}v$ is found to be convenient and efficient, requiring only a single matrix multiply.

In working with stochastic data recovery, it is obvious that $B_{mn}v$ using the Bernstein functions mollifies the data vector $v$, and thus provides an approximation vector of length $m$ to the initial vector $v$ of length $n$. Consider the evaluation of $B_{mn}A^{-1}_{nn}v$ where the generator of the row space of $A_{nn}$ and $B_{mn}$ are given by (2). This interpolates the data provided that $\sigma_B$, meaning that the variance of the Gaussian pdf that is used to generate $B_{mn}$, is the same as the variance $\sigma_A$ that is used to generate $A_{nn}$. If instead $\sigma_B > \sigma_A$, then the preimage $A^{-1}_{nn}v$ will be oversmoothed when $B_{mn}$ is applied to the preimage, and the result will be approximation. Similarly, if $\sigma_B < \sigma_A$, then the preimage $A^{-1}_{nn}v$ will not be smoothed sufficiently, and the data will be roughened, or more appropriately it will be deconvolved.

With statistical errors present in multidimensional data, interpolation in parametric coordinates may yield an extremely complex curve, and the errors may cause the curve to wiggle excessively, often crossing over on itself. Clearly, some form of smoothing is necessary. However, as noted in Section 1, simply approximating the data so as to smooth these errors may introduce translation errors in the approximating surface representation, particularly when the number of data points is small, or the smoothing specified by $\sigma$ is large. Since the approximation is convergent, a simple work-around can be achieved by densification of the data by interpolation as this will minimize this shift on subsequent smoothing. This leads to a computationally efficient approach to surface recovery that avoids translation errors while smoothing the noise based on evaluating $B_{mn}A^I_{nn}v$, where $m$ is the desired number of output points representing the boundary, and $n$ are the number of input data points and $\sigma_B > \sigma_A$.

To demonstrate this form, note that the intended construction is to evaluate $B_{pn}A^{-1}_{nn}v$, where $p$ is significantly greater than $n$ with $\sigma_B = \sigma_A$, and then applying smoothing to this densified data by multiplying by $C_{mp}$, where $\sigma_C > \sigma_A$. In effect, this interpolates the data and then smooths it by the application of the approximation $C_{mp}$ to the densified data. Thus a two-step algorithm can be described as follows:

```
(1) compute densified interpolant $B_{pn}A^{-1}_{nn}v$ to
    the data vector $v$;
(2) compute boundry approximant $C_{mp}(B_{pn}A^{-1}_{nn}v)$.
```

This algorithm is equivalent to the following:

```
(1) compute densified boundry approximate
    interpolant $B_{mn}A^I_{nn}v$, to the data vector $v$;
```

that is, there exist $\sigma_C$ and $\sigma_B$ such that applying $C_{mp}$ to $B_{pn}A^{-1}_{nn}$ is the same, or nearly the same, as applying $B_{mn}$ to $A^I_{nn}$. The use of $A^I_{nn}$ for a wide range of $\sigma_A$ instead of $A^{-1}_{nn}$ introduces some additional smoothing, allowing for less smoothing to be used on the convolution step, that is, when applying $B_{mn}$. Since it saves an unnecessary matrix multiply, it is clearly faster. The construction of $A^I_{nn}$ is not difficult, remarkably being given by the inverse of the generator of the row space of $A_{nn}$. The elements of the approximate inverse are given by the reciprocals of the coefficients of $A_{nn}$. It is for this reason that the direct inversion of $A_{nn}$, or solution of the system using efficient Toeplitz solvers, is not needed. The only exception may be when the data is free of noise and exact interpolation without any smoothing is desired.

The reconstruction of surface data is done using a parametric representation $\{s_i\}^n_{i=1}$ in which $s_i = (x(t_i), y(t_i))$ for two-dimensional data, and $s_i = (x(t_i), y(t_i), z(t_i))$ for three-dimensional data, where $0 \leq t_i \leq 1$. The recovery of the data is done using (1) or its modification using an approximate inverse $A^I_{nn}$, applied successively to the data pairs $\{(t_i, x_i)\}$, $\{(t_i, y_i)\}$ in two dimensions and to the data triplets $\{(t_i, x_i)\}$, $\{(t_i, y_i)\}$, $\{(t_i, z_i)\}$ in three dimensions. For example, in the case of interpolating two-dimensional data, we apply $B_{mn}A^{-1}_{nn}x$ and $B_{mn}A^{-1}_{nn}y$ to obtain the interpolant to the position vector $x$ and the position vector $y$, or we apply $B_{mn}A^I_{nn}x$ and $B_{mn}A^I_{nn}y$ to obtain the approximate interpolants.

In reconstructing a parametrically represented surface generated from image data from pixel values, for instance, an algorithm for boundary detection and sorting is necessary. In our analysis, it is assumed that this is available, however the errors generated in parameterizations of the surface may not be entirely random, and thus systematic errors in surface representation will also be introduced. For the purpose of assessing the performance of the boundary recovery algorithms, it will be assumed that the errors are random Gaussian with a mean of zero, with variable variances $v$, generated using the random variable

$$\xi = v\sqrt{-2\log(r_1)}\cos(2\pi r_2), \tag{7}$$

where $r_1$ and $r_2$ are two random numbers uniformly distributed in the interval $[0,1]$. Thus the parametrically ordered data $\{(x_1(t_i), \ldots, x_d(t_i))\}$, with $d = 2$ or 3, and $i = 1, \ldots, n$, are perturbed to yield the data sets $\{(x_1(t_i) + (\xi_1)_i, \ldots, x_d(t_i) + (\xi_d)_i)\}$.

In applying stochastic data recovery to the problem of finding the shape of a parametrically defined boundary, the problem of closed curves needs to be addressed. In the presence of large amount of noise, the two endpoints of the parametrically defined curves may not match: while ideally $(x_1(t_1), \ldots, x_d(t_1)) = (x_1(t_n), \ldots, x_d(t_n))$ in the case of a close loop, in the presence of errors this will not be the case.

Finally, in implementing the algorithm, it was found that the dependence of (2) on $\sqrt{n}$ in the denominator made the choice of $\sigma$ dependent on $n$, as the boundary data density increased, the recovery using any given value of $\sigma$ produced increasingly rougher curves as $n$ increased, and thus it was found convenient to evaluate $a_{jk}$ and $b_{jk}$ based on a constant value of $\sigma$. Additionally, the algorithm was applied to all of the boundary data associated with a parametric data set to construct the boundary curve, rather than decomposing the data into overlapping blocks. The merits of using all of the data are a slight improvement in accuracy, while the drawbacks are that the cost of evaluating the algorithm increases as the block size increases, and this should be born

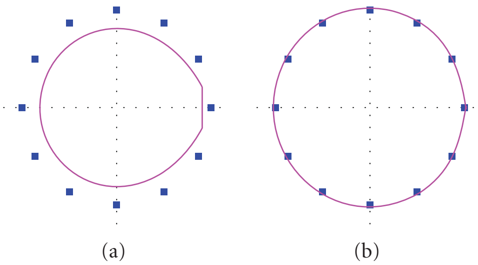


Figure 1: Recovery of the boundary of disk specified by 12 points, starting at $(1, 0)$ on the $x$-axis, and moving counterclockwise around the circle. Note that in the circle in (a) the boundary of the recovered disk is approximated with $\sigma_B = 0.001$ using $B_{100,12}v$, resulting in the inscribed disk, while in (b) the boundary is constructed using approximate interpolation using $\sigma_A = \sigma_B = 0.001$ and applying $B_{100,12}A^I_{12,12}v$.

in mind in applying the algorithm to large complex three-dimensional data sets.

The choice of $\sigma$ depends on the smoothness of the desired boundary curve. The larger is the value of $\sigma$, the smoother are the results. The values for $\sigma_A$ and $\sigma_B$ depend on the amount of noise, as well as on the presumed smoothness of the boundary data. While this seems to present difficult choices, it is less complicated than it appears, as the choice for $\sigma_A$ will usually be any sufficiently small value. For example, setting $\sigma_A \leq 10^{-6}$ allows for representing the data as nearly piecewise linear along the boundaries. In contrast, the choice for $\sigma_B$ requires some evaluation as this determines the smoothness of the recovered boundary.

## 3. RESULTS AND DISCUSSION

We begin by applying the process to the recovery of the boundary of a disk defined parametrically by 12 uniformly distributed points with no random errors in the data as shown in Figure 1. The figure clearly illustrates the aforementioned difficulty of attempting to approximate (smooth parametric data).

In examining Figure 1(a), it is also important to observe that the beginning and the end of the curve are joined by a straight line in this example. The reason is that the slope of the approximant to the data $(x(t_1), y(t_1))$, $(x(t_2), y(t_2)), \ldots$ is not the same as the slope to the data $\ldots, (x(t_{n-1}), y(t_{n-1})), (x(t_n), y(t_n))$, and thus in this example, the first approximated point $(r_1, s1)$ does not agree with the last approximated point $(r_m, s_m)$, and are joined graphically with a straight line to close the curve.

A solution to the mismatch is to overlap the curves during reconstruction, that is, at several points away from the last point, and ending the construction several points away from the first point, then only using the curve from the first to the last point. In all of the studies presented, there is no attempt to overlap the curves in order to magnify these boundary affects, and to demonstrate that they are mostly negligible, as seen in Figure 1(b), whenever the construction is done correctly. For large data sets where it may be computationally advantageous to block the data, overlapping the endpoints is readily accomplished.

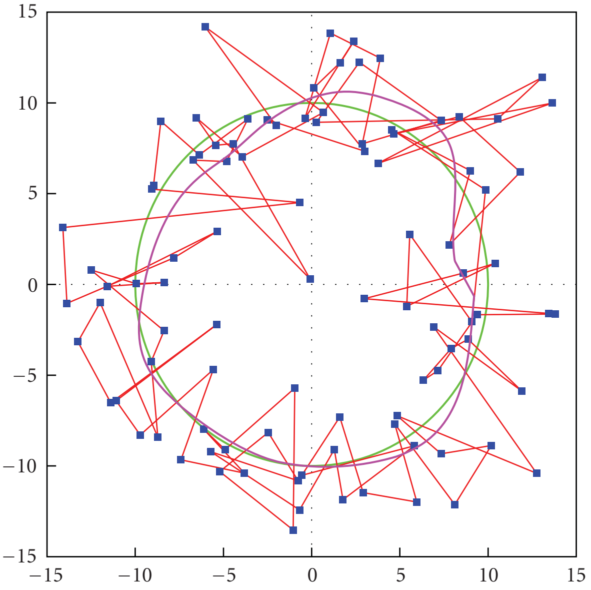


Figure 2: Recovery of the boundary of disk specified by 96 points with Gaussian noise of $\nu = 3$, starting at $(1, 0)$ on the $x$-axis and moving counterclockwise around the circle with the boundary recovered using approximate interpolation, with $\sigma_A = 1 \times 10^{-6}$ and $\sigma_B = 1 \times 10^{-4}$. The recovered curve is shown as a thicker line.

It is important to realize that if the number of points representing the disk were to be much more than 12, then it would have been difficult to visualize the contraction of the recovered boundary curve since the approximation is convergent. Thus for a sufficiently large number of data points, the difference between the approximating curve and the curve itself becomes arbitrarily small.

Recovering the boundary of a disk becomes more difficult, as shown in Figure 2, particularly if the amount of noise is quite large. In this example, the Gaussian noise for a disk of radius 5 is specified as $\nu = 3$ yielding an extensively scattered data set. While the level of noise in this illustrative example is much higher than would be expected in any realistic imaging situation, the example serves a two-fold purpose: (1) on the one had it shows the robustness of the method at recovering a reasonable representation of the surface from the data that is more consistent with noise than data, and (2) it shows that the effects of errors in constructing the parameterization are less of an issue than might be presumed. In the figure, the connectivity of the boundary data is not ordered in $\theta$ moving counterclockwise around the circle, and so it is quite likely that any minor errors in parameterization, for example, using even the simplest unconstrained nearest neighbor search of the data, would cause unrecoverable errors in the surface representation that is recovered.

As expected, reducing the noise to $\nu = 1/2$ significantly improves the recovery, even for half as many points, as shown in Figure 3.

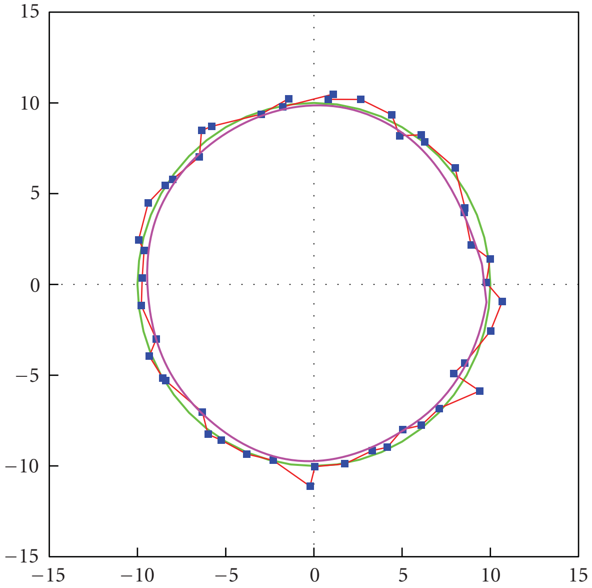


Figure 3: Recovery of the boundary of disk specified by 48 points with Gaussian noise of $\nu = 1/2$ as in Figure 2.

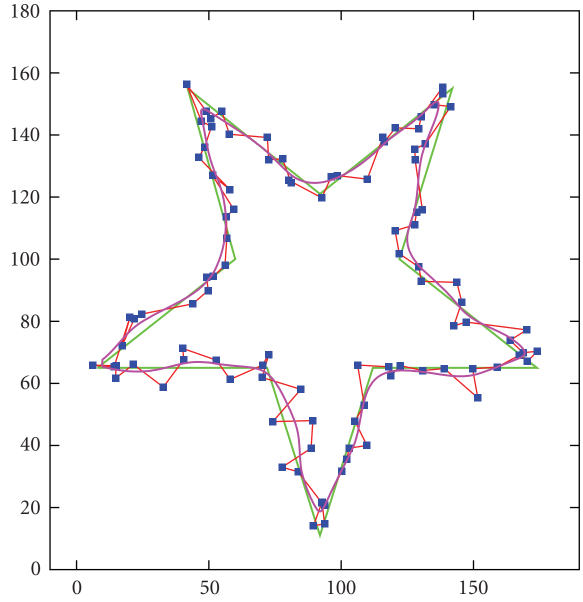


Figure 4: Recovery of the boundary of the star specified by 100 points: 10 along each arm with Gaussian noise of $\nu = 4$. The boundary is constructed using approximate interpolation, using $\sigma_A = 1 \times 10^{-6}$ and $\sigma_B = 1 \times 10^{-4}$. The recovered curve is shown as a thicker line.

While both of these figures are typical, the symmetry and the smoothness of the boundary of the disk make the recovery somewhat less challenging than for a more complex geometrical shape. Thus the performance of the algorithm is examined on a star-shaped region generated from the vertex data {(8,65), (72,65), (92,11), (112,65), (174,65), (122,100), (142,155), (92,121), (42,155), (60,100), (8,65)}. Note that the recovery makes use of $\sigma$ that on both steps is the same as that used in recovering the boundary of the disk.

The shape of the star in Figure 5 has become more wiggly using the smoothing parameters the same as in the case of lesser noise. The problem is a classical one: there is no means to discern the shape of the figure from the noisy data, except to note that an acceptable shape is determined by the smoothness of the boundary that is intrinsic to the figure. In this case, changing the smoothing can accommodate this subjective assessment, as illustrated in Figure 6. Note that the recovered boundary is consistent with the curve recovered from the less noisy data set: compare Figures 4 and 6 as shown in Figure 7.

The effects of doubling the smoothing by taking $\sigma_B$ to be twice as large are clearly evident. Since the noise in both cases was generated using the same seed, it is only the magnitude of the excursions away from the star's boundary that change, and hence the figures are directly comparable. This perhaps most plainly illustrates that interaction between noise and smoothing, as the two curves are nearly identical. Even when the noise is doubled again to $\nu = 16$, the shape of the recovered curve using $\sigma = 0.0004$ is remarkably consistent, as shown in Figure 8. At this level of noise there is some loss of resolution of the limbs, however the recovered boundary is recognizable as being related to the two boundaries obtained in Figures 4 and 6. While this is artificial in that the amount of noise in the data in real problems is not known, it does demonstrate the robustness of the algorithm at consistently recovering the boundary data irrespective of the added noise.

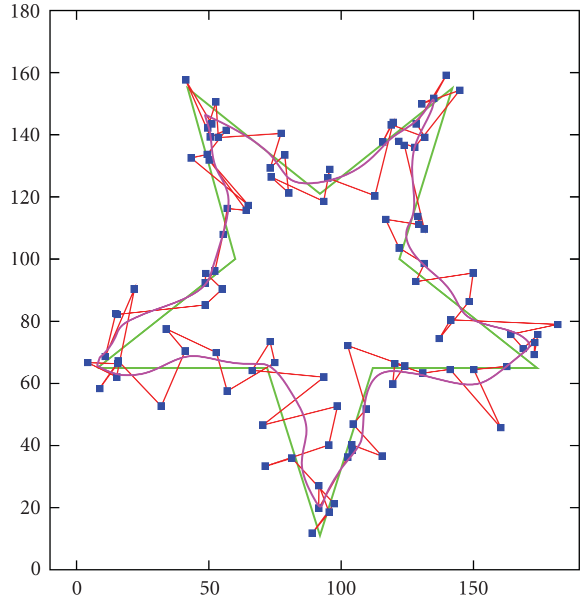


Figure 5: Recovery of the boundary of the star specified by 100 points with Gaussian noise of $\nu = 8$ as in Figure 4. The recovered curve is shown as a thicker line.

A final remark on the computing of the approximate interpolant is the following. Since the approximate interpolant fails to interpolate when $\sigma$ is large or, more appropriately, fails to interpolate rapidly varying data, quite some effort was expended in developing mechanisms for

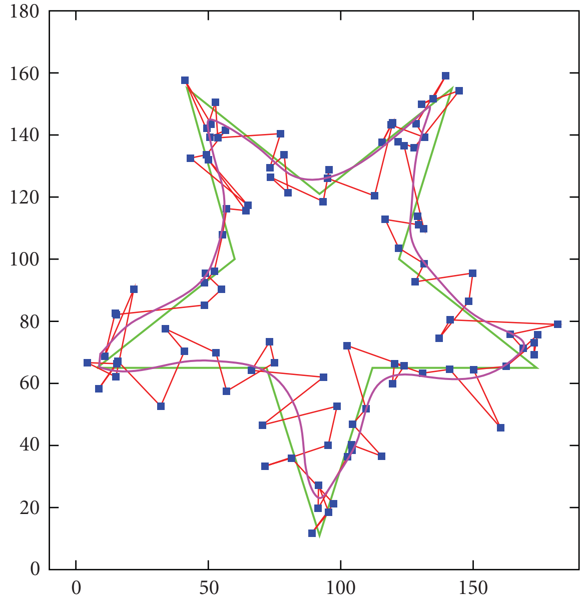


Figure 6: Recovery of the boundary of the star specified by 100 points with Gaussian noise of $\nu = 8$ as in Figure 5. The boundary is constructed using approximate interpolation with $\sigma_A = 1 \times 10^{-6}$ and $\sigma_B = 2 \times 10^{-4}$. The recovered curve is shown as a thicker line.

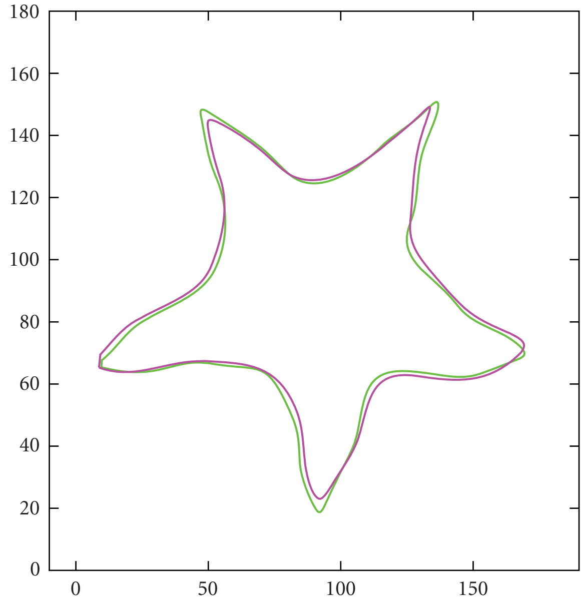


Figure 7: Recovery of the boundary of the star comparing the recovered boundary shown in Figure 4 to the recovered boundary shown in Figure 6. In this comparison, the recovery algorithm is implemented so that the smoothing is proportional to the noise in the cases being compared, yielding excellent results.

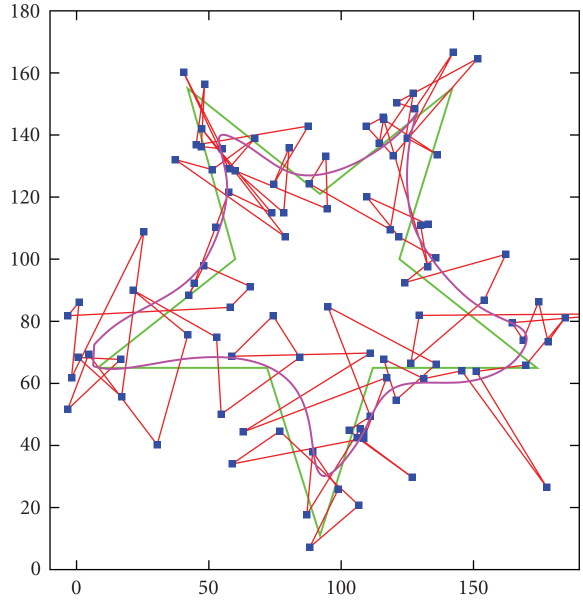


Figure 8: Recovery of the boundary of the star specified by 100 points with Gaussian noise of $\nu = 16$. In this case, an extremely large amount of noise is added, however the algorithm is effective at recovering features of the star shape. The boundary is constructed using approximate interpolation by using $\sigma_A = 1 \times 10^{-6}$ and $\sigma_B = 4 \times 10^{-4}$. The recovered surface is shown as a thicker line.

being able to use large $\sigma$ for smoothing and yet still maintain some fidelity with the boundary data while constructing the preimage. As it developed, this iterative correction was not necessary: the algorithm performed well even without the introduction of any corrections. In part, this is due to the rather simple shapes, and the relatively large amounts of noise that were examined.

In the case when a surface is oscillating rapidly with the noise much less than this surface oscillation, it is clearly necessary to implement the algorithm using this additional correction. Indeed, given sufficiently rough surface data, it may be necessary to use stochastic interpolation as otherwise some high-frequency surface details will be oversmoothed by the approximate inverse construction.

## 4. CONCLUSIONS

Examination of several attempts at domain boundary reconstruction in two dimensions is encouraging using the stochastic data recovery techniques. In particular, qualitative assessments demonstrate the viability of utilizing stochastic function recovery methods for reconstructing parametrically defined edges. Since the approach is intrinsically one dimensional, its extension to three dimensions is not difficult, and thus can easily be implemented and tested on more realistic problems. Moving to three dimensions poses no additional cost other than that more data has to be processed, that is, the algorithmic costs scale directly with the number of lines being evaluated, and the number of points on each line at which the algorithm is applied.

It should be noted that the proposed technique does not solve all problems involving noisy surface reconstruction, however it does provide an additional tool for analysis. The Gaussian-based kernels (the Bernstein functions) which were used to generate the row space of the matrices, were remarkably effective at cancelling the noise even under the most extreme conditions where the noise essentially obliterated the image shape. Of course, this noise was Gaussian and so it is only reasonable that the proposed approach would work well in these circumstances. For other types of noise, the use of alternative probability density functions for generating the mollifiers is easily accomplished, and thus the method has substantial design flexibility, and these options need to be explored in detail to ascertain their utility at cancelling these other types of noise.

The computational advantages of the technique are that it requires only two matrix multiplies of the data vector, and thus the approach is relatively cost effective. Moreover, the method is easily implemented in parallel, and further computational gains in efficiency can be achieved by blocking the data since typically only a segment of the entire data vector is needed to recover the data in any region. If fixed block sizes can be implemented, then the cost of the two matrix vector multiplies can be further reduced as the matrix-matrix multiply needs to be done only once, and so the algorithm reduces to a single-block matrix-vector multiply.

## REFERENCES

[1] J. Kolibal and D. Howard, "The novel stochastic Bernstein method of functional approximation," in *Proceedings of the 1st NASA/ESA Conference on Adaptive Hardware and Systems (AHS '06)*, pp. 97–100, Istanbul, Turkey, June 2006.

[2] I. Kastanis, S. R. Arridge, A. M. S. Silver, D. L. Hill, and R. Ravazi, "Reconstruction of the heart boundary from undersampled cardiac MRI using Fourier shape descriptors and local basis functions," in *Proceedings of the 2nd IEEE International Symposium on Biomedical Imaging: Macro to Nano (ISBI '04)*, vol. 2, pp. 1063–1066, Arlington, Va, USA, April 2004.

[3] Z. Malchano, "Image guidance in cardiac electrophysiology," Massachusetts Institute of Technology, Cambridge, Mass, USA, 2006.

[4] J. Xie, Y. Jiang, and H.-T. Tsui, "Segmentation of kidney from ultrasound images based on texture and shape priors," *IEEE Transactions on Medical Imaging*, vol. 24, no. 1, pp. 45–57, 2005.

[5] D. Howard and J. Kolibal, "Image analysis by means of the stochastic matrix method of function recovery," in *ECSIS Symposium on Bio-Inspired, Learning, and Intelligent Systems for Security (BLISS '07)*, pp. 97–101, Edinburgh, UK, August 2007.

[6] J. Kolibal and D. Howard, "Implications of a novel family of stochastic methods for function recovery," in *Proceedings of the 2nd International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '06)*, pp. 495–498, Pasadena, Calif, USA, December 2006.

[7] J. Kolibal and D. Howard, "MALDI-TOF baseline drift removal using stochastic bernstein approximation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 63582, 9 pages, 2006.

[8] W. F. Trench, "An algorithm for the inversion of finite Toeplitz matrices," *SIAM Journal on Applied Mathematics*, vol. 12, no. 3, pp. 515–522, 1964.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 259174, 10 pages
doi:10.1155/2008/259174

*Research Article*

# Bio-Inspired Microsystem for Robust Genetic Assay Recognition

**Jaw-Chyng Lue[1] and Wai-Chi Fang[2]**

[1] *Department of Electrical Engineering - Electrophysics, University of Southern California, Los Angeles, CA 90089, USA*
[2] *Department of Electronics Egineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 300, China*

Correspondence should be addressed to Jaw-Chyng Lue, lormen@gmail.com

Received 14 January 2008; Accepted 25 March 2008

Recommended by Daniel Howard

A compact integrated system-on-chip (SoC) architecture solution for robust, real-time, and on-site genetic analysis has been proposed. This microsystem solution is noise-tolerable and suitable for analyzing the weak fluorescence patterns from a PCR prepared dual-labeled DNA microchip assay. In the architecture, a preceding VLSI differential logarithm microchip is designed for effectively computing the logarithm of the normalized input fluorescence signals. A posterior VLSI artificial neural network (ANN) processor chip is used for analyzing the processed signals from the differential logarithm stage. A single-channel logarithmic circuit was fabricated and characterized. A prototype ANN chip with unsupervised winner-take-all (WTA) function was designed, fabricated, and tested. An ANN learning algorithm using a novel sigmoid-logarithmic transfer function based on the supervised backpropagation (BP) algorithm is proposed for robustly recognizing low-intensity patterns. Our results show that the trained new ANN can recognize low-fluorescence patterns better than an ANN using the conventional sigmoid function.

## 1. INTRODUCTION

The development of low-cost portable instruments for rapidly analyzing genetic assays would significantly advance the level of medical services globally. The polymerase chain reaction (PCR) amplification and the capillary electrophoretic (CE) techniques are often adopted for genetic analysis. A complex system that can process full PCR amplification and data analysis tasks usually involves integration of control, optical, thermal, fluid channel, and data acquisition systems. For example, a portable system providing full PCR-CE functions was developed earlier for genetic analysis [1]. The system demonstrated the feasibility of on-site genetic analysis. However, the expense to build such a system is considered relatively high. The entire integrated system consists of multiple PCR chambers, heaters, sensors, solid-state laser excitation light source, fluorescence detection optics, electronics, CE separation microchannels, and power supplies. The data was collected and processed in a portable computer. Recently, a real low-cost (∼10US$) pocket-sized PCR thermocycling device has been developed based on a smart technique of simultaneously pseudoisothermally heating multiple zones of a loop channel for PCR amplification [2]. This thermocycler does not contain the CE separation, the fluorescence detection, and the data analysis functions. Theoretically, multiple PCR amplification results can be rapidly generated in parallel and displayed simultaneously by using multiple of these low-cost devices. Therefore, patterns of an array of the PCR resulted samples can potentially be generated similar to the genetic assay patterns on a microchip. Moreover, the integration of PCR and electrochemical (EC) transduction functionality on microfabricated silicon/glass-based devices for DNA amplification and detection was shown successfully [3]. Their microfabricated device needs to operate with external control and data-acquisition systems.

Most of the research efforts for PCR analysis tools were focusing on the development of the PCR microdevices, the associated thermal systems, the optical systems, and the data analysis software tools. However, to our best knowledge, the data acquisition and analysis system for examining PCR samples or assays is usually a computer equipped with specific PCR analysis software but not a compact hardware solution.

Regarding the goal of building a real compact PCR analysis system that can rapidly find and analyze the desired genetic patterns, the existing data acquisition and analysis systems (e.g., portable computers and interfaces) are considered relatively large in size and heavy in weight. In addition, human inspectors cannot recognize the genetic assay patterns as easily as written characters with explicit meanings. Manual massive PCR data analysis can be very time consuming. Therefore, people involved in "the human genome project" have used perceptron-like neural networks for helping to recognize the DNA fragments with specific functions [4]. For gene-recognition purpose, a perceptron was first trained by using the datasets consisting of nucleotide sequences of known functional sites (e.g., transcription initiation sites (promoters), transcription termination sites (terminators), or splice-junction sites). Patterns of fragments of the entire DNA sequence were then fed to the perceptron nucleotide by nucleotide to check if any site of interest appears at a particular position in the fragment. In the protein coding region recognition task by Guan et al. [5], their multisensor /neural network successfully identified 96% of the 17,576 sequence positions as coming from coding or noncoding regions. In the splice junction recognition experiment, the resulting recognition rate was 99% for acceptor junctions and 96% for donor junctions.

Nonetheless, if the genetic analysis task needs to be conducted on a hazardous or dangerous field (e.g., potentially disease-contagious environment), a compact, autonomous, and even disposable PCR data analysis system would be preferred. Therefore, by taking advantages of the VLSI microfabrication technologies and artificial neural network theories, we proposed a microsystem consisting of a unique optical configuration setup, a differential logarithm sensor-processor array chip, and an ANN SoC processor chip for fast recognizing and analyzing the PCR prepared genetic patterns.

In typical PCR amplification procedure, a dual-labeled (i.e., for sample and reference channels) assay design is commonly used for identifying differentially expressed genes. This method also reduces the sources of variability/noise due to aspects of individual spot that affects both specimens similarly [6]. In order to accurately calculate the density of the sample DNA material in a particular dot/well after the PCR amplification, the integral of the total fluorescence intensity (presumably representing the density of the DNA materials inside the dot/well) from the topological profile of the dot/well is usually computed. The logarithmic value of the ratio of the two intensities of the fluorescent-dye-labeled specimens (one for the sample specimen, the other for the reference specimen) measured from the same dot/well is calculated based on the fluorescence assay image. The ratio of the two intensities would provide the normalized population of the genetic material in the dot/well disregarding the initial population density before PCR amplification. The logarithmic operation would amplify the small signals. In most of the commercial available solutions, the fluorescence assay image is usually scanned by a color scanner with high resolution and then transferred to a computer for image analysis. The profile analysis software usually computes the normalized intensity of dot/well after dot/well sequentially.

The intensity of fluorescence light is usually relatively low. Using higher excitation light intensity can lead to brighter fluorescence patterns. Increasing the integral of detection time can enhance the received fluorescence patterns. However, lower power consumption and faster detection are preferred. Furthermore, some fixed-pattern noises in the input pattern may exist (e.g., fixed pattern noises created by scattered lights, nonuniformity of the responsivity of the detector array). These noises may introduce errors to the measurement of the density of the DNA materials.

In order to fast parallel-process the data and resolve the ambiguity induced by the noises in the data analysis task, a trained artificial neural network is considered a solution. The parallel processing capability comes from the nature of the ANN's multiple input architecture [7–9]. In contrast with the conventional sequential data analysis methods, the data analysis throughput would be increased linearly as the number of dot/well increases. Regarding potential noises, the ANN will automatically take the noises into calculation in the latter recognition phase because the ANN can learn from the training patterns that contain the fixed pattern noises in the learning phase. In addition, because of the natural capability of associating noisy input patterns with output index of a trained ANN, noises introduced by other factors will not significantly affect the ANN's recognition capability. In conjunction with the natural capabilities of an ANN listed above, a signal amplification stage that can augment the low-fluorescence input before the ANN stage would help the ANN to acquire data more reliably, and thus result in a more robust data analysis capability of the entire system.

## 2. BIOCHIP MODULE ARCHITECTURE

We proposed a hardware microsystem that is suitable for real-time, on-site, robust genetic fluorescence data analysis (Figure 1). This envisioned biochip module architecture consists of an on-chip assay with an array of clusters of dual-labeled genetic dots/wells, a dual-color beams module, an imaging lens, a bioimaging optoelectronic microchip with coated color filters (Figure 2), a parallel analog data-transfer bus (optional depending on the implementation method), and an artificial neural network (ANN) module for image analysis.

The operational function of each module is explained below along the optical and electrical signal pathways. The dual-labeled genetic dots/wells are simultaneously excited by two monocromatic excitation beams (e.g., 532 nm with a bandwidth of 10 nm from a green diode laser pointer source for the cyanine Cy3 dye, and 635 nm with a bandwidth of 10 nm from a red solid-state diode laser source for the cyanine Cy5 dye) according to the receiving bandwidths of the sample and the reference channels. The assay can be either front-side or backside illuminated as long as a clear fluorescence image of the dot/well array is generated. Two fluorescence patterns with different peak wavelengths are produced (e.g., peak value at 570 nm from the Cy3 dye and peak value at 670 nm from the Cy5 dye, the two

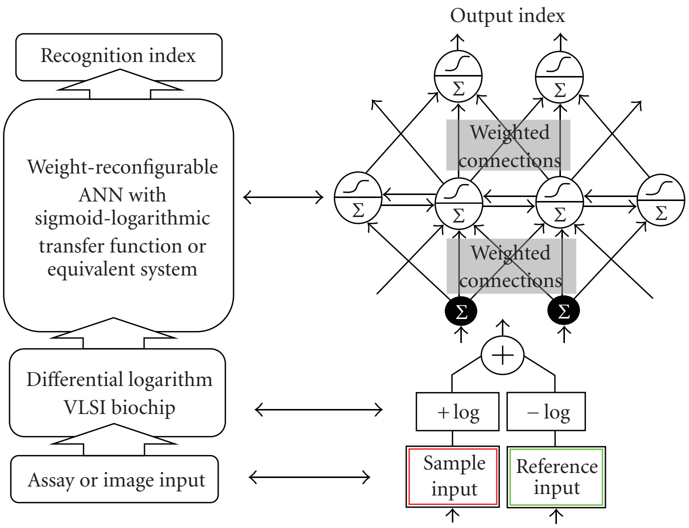


Figure 1: (a) Hierarchical diagram of the proposed biochip microsystem for genetic assay recognition. (b) Schematics of a three-layered ANN and the preceding differential logarithm stage. The system of dual-labeled gene assay, dual-color beam module, and imaging lens is not shown in this schematic diagram. The thin-film color filters coated on top of the sample and reference channels are represented by the red and green boxes.

spectral profiles are highly distinguishable), and imaged onto the bioimaging chip through an imaging lens. Each unit on the bioimaging chip contains two sensor channels. One sensor is coated with a thin-film microfilter for wavelength A (e.g., 580 nm with a narrow transmission bandwidth of approximately 40 nm of a deposited thin-film filter [10]). The other sensor is coated with a thin-film micro-filter for wavelength B (e.g., 675 nm with a transmission bandwidth of 40 nm). The two optical signals are received simultaneously and processed separately through the posterior electrical circuits. A bioimaging chip made of an array of differential logarithm circuitry was designed (Figure 2 (bottom left)). Two analog photoelectric voltage signals produced by the separated logarithmic amplifier circuit channels are fed continuously into the differential pair circuit. The difference of the two input voltages is then represented by the output voltage from the differential pair circuit. The calculation of the logarithm of the ratio of the sample to the reference fluorescence intensities is effectively accomplished in this chip. The analog output from each unit in the bioimaging chip is then sent to the posterior artificial neural network module through the parallel data bus.

The ANN stage is responsible for filtering and recognizing the desired assay cluster patterns. Fixed pattern noises and noises caused by the nonlinear circuitry are expected to be accommodated after the ANN is trained. Either unsupervised or supervised learning algorithm can be adopted to train the ANN. The ANN in the biochip module architecture can be implemented by either hardware or software. In this work, we provide a hardware implementation (i.e., a weight-reconfigurable winner-take-all ANN chip suitable for the Kohonen self-organized filter algorithm [7]) for the unsupervised version, and a computer-simulated ANN (i.e., a feedforward ANN using the back-propagation (BP) learning algorithm [8, 9]) for the supervised version. In the hardware implementation, the weight values are reconfigurable and stored in memory devices. By adopting a massively paralleled neural computing paradigm and a mixed-signal deep submicro fabrication technology, the ANN can be implemented on a single VLSI chip. A row/column parallel data flow architecture is used to connect all on-chip systems, and to reduce data bandwidth limitations due to conventional data bus architectures.

Because the weak fluorescence signals are enhanced by the imaging chip and automatically analyzed by the noise-tolerable neural network module, the entire architecture system is expected to robustly conduct the recognition task.

## 3. HARDWARE MODULE IMPLEMENTATION

### 3.1. Bioimaging chip and its nonlinear circuitry

The proposed bioimaging chip consists of an array of differential logarithm processor unit and row/column readout circuit. A prototype layout of an array of 15 × 15 differential logarithm unit is shown in Figure 2. Each unit contains two size-matched logarithmic amplifier circuits and a differential pair circuit. As described in the previous section, each unit produces an analog voltage output to represent the difference between the logarithms of the sample (experimental) input and the reference (control) input.

The key logarithmic amplifier circuit is designed after the works of Chamberlain and Lee [11] and Mead [12], but with an additional $n$-well/$p$-sub junction layer for effectively isolating cross-talk noises among the logarithmic amplifier processor unit array.Chamberlain and Lee first adopted the intrinsic vertical $n$-$p$ junction ($n^+$-diffusion/$p$-sub) under the source terminal of an NPN transistor to

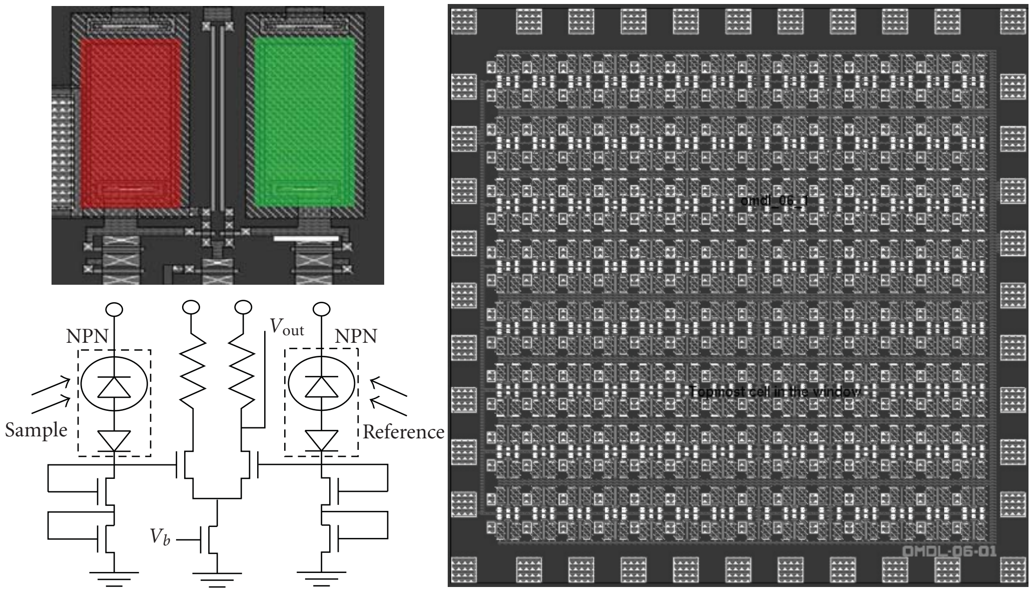


Figure 2: (right) Proposed layout of a 15-by-15 unit cell array of the differential logarithm circuitry (2.2 mm by 2.2 mm), (top left) an enlarged view of the layout of a single cell with pseudo thin-film monochromatic filter layers, and (bottom left) the schematic diagram of a single unit cell.

function as a reverse-biased photodiode. A wide dynamic range silicon photodetector can be implemented [11]. Mead has a similar design by using a vertical parasitic bipolar transistor for photosensing part [12]. This vertical bipolar is a natural byproduct structure in the CMOS process. The $p^+$-diffusion/$n$-well/$p$-sub structure is the PNP bipolar transistor existing in any PMOS transistor in an isolated section of $n$-well region. This design is the fundamental building block of his silicon retina.

The optoelectronic logarithmic amplifier circuit for each channel in this work was fabricated by using the MOSIS AMI 1.5-$\mu$m 5-V ABN BiCMOS $n$-well process as shown in Figure 3 [13]. The photodetector is made of a vertical $n$-well/$p$-base/$n^+$-emission bipolar detector. Two diode-connected NMOS transistors are connected in series with the bipolar detector. The detecting area collects the fluorescence inputs. Mainly due to the diode configured NMOS transistors, the V-I characteristic of this normally “OFF” and low-power circuit behaves logarithmically while operating in the subthreshold region, and similar to a square root curve ($\sim\sqrt{I_{\text{DS}}}$) while operating in the saturation region. This logarithmic amplifier circuit can detect light intensity as low as approximately 10 nW and consumes energy from 100 nW to 2 $\mu$W ($V_{\text{DD}}$: 5 Volts) depending on the incident intensity and wavelength.

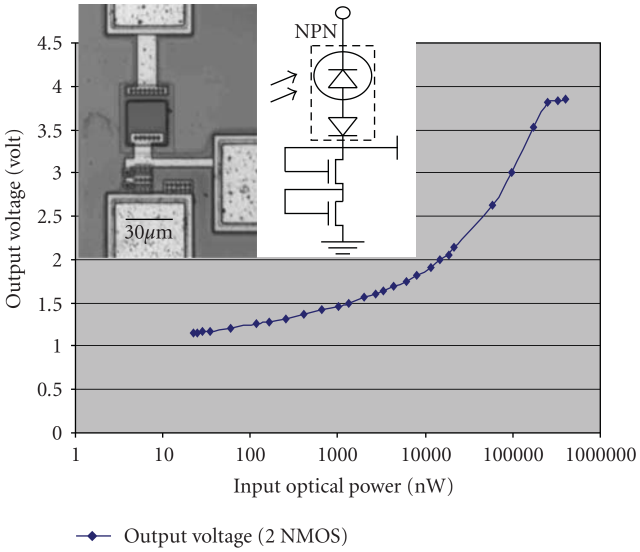


Figure 3: Output voltage of a single channel of saturated logarithmic circuit as function of the input optical power (input wavelength: 830 nm). An optical micrograph of the single-channel logarithmic amplifier circuit and its correspondent schematic circuit diagram are shown.

### 3.2. Weight-reconfigurable artificial neural processor chip

The SoC architecture design of the weight-reconfigurable ANN processor consists of an input neurons array, a programmable synapse weight matrix, an array of output neurons, a winner-take-all module, and a membership encoder [14, 15]. The input neurons array has $M$ input neurons that are used to buffer the input vector. Each input vector has $M$ analog components (generating from the preceding bioimaging chip). The programmable synapse weight matrix is composed of $M \times N$ synapse cells for the $NM$-dimensional codevectors. The output neuron array is composed of $N$ summing neurons with selectable sigmoid or sigmoid-logarithmic transfer functions. The winner-take-

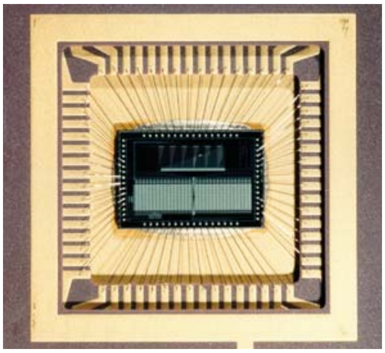


Figure 4: The optical micrograph of the prototype ANN chip that is wire-bonded to a ceramic package. The silicon chip die size is 4.6 mm × 6.8 mm.

all module consists of $N$ competitive circuit cells, which perform parallel comparison among $N$ inverted distortion values and choose a single winner. The membership encoder circuit is an $N$-to-$n$ decoder that uses binary codes to encode $N$ classes.

The ANN processor works as a learning accelerator in the learning phase at a time complexity $O(1)$ for processing each learning iteration. Its programmable weight matrix can be either generated by using the on-chip self-organization learning procedure or be uploaded by the BP training subsystem [15]. The ANN processor also realizes a full-search vector quantization process for each input vector at a time complexity $O(1)$ in the recognition phase.

This ANN processor can also support the multiple winner-take-all scenario (e.g., more than one classes that the input assay pattern may belong to, or multiple desired patterns that the input assay pattern are similar to). After a winning pattern (the most likelihood) was picked out from the $N$ prestored classes/codevectors in the recognition task according to a particular analog input vector, the associated circuitry of this winning pattern will be disabled in the next recognition iteration. Therefore, a second-winner pattern (the second likelihood) can be chosen later according to the same input vector. By repeating the procedure stated above, multiple-winner patterns could be chosen for the current input pattern eventually.

The ANN chip can learn unsupervised if the selforganization learning procedure is adopted. In this case, the ANN chip can perform on-chip learning in the learning phase. For the supervised learning version (e.g., back-propagation algorithm or its variations), the weight update procedure usually involves complex computations that require further signal processing circuits in order to achieve the on-chip learning purpose. Further real estates on chip are then required to accommodate the circuits.

A prototype ANN SoC chip using a scalable 2-$\mu$m 5-V CMOS technology was designed, fabricated, and tested. Its chip layout and design features are shown in Figures 4 and 5, respectively. This prototype chip includes 25 input neurons, 25 × 64 weight cells, 64 output summing neurons, 64 winner-take-all cells, and a 64-to-8 membership encoder. The estimated power dissipation is 50 mW at 10 MHz. Its equivalent computation power is about 16 giga-operations per second.

An engineering version of the ANN SoC SiP (silicon intellectual property) has been under development to enable the proposed miniaturized PCR system-on-chip design using the TSMC 130-nm 1.2-V CMOS technology. The scalable ANN prototype chip can be converted into a design containing 100 input neurons, 100 × 256 weight cells, 256 output summing neurons, 256 WTA cells, and a 256-to-8 membership decoder. The envisioned chip size is approximately 1.2 mm × 2 mm. Its estimated power dissipation is about 120 mW at a 100D vector throughput rate of 100 MHz. Its equivalent computation power is about 2.5 tera-operations per second. Because of this ANN SoC SiP design, the feasibility of the proposed low-power, real-time, and on-site PCR assay analysis on an integrated microsystem becomes promising. The proposed microsystem would be useful especially in a scenario of finding desired or suspicious biopatterns in a massive amount of data.

## 4. SIMULATIONS AND EMPIRICAL RESULTS

### 4.1. Numerical simulations of biosignature and optical character recognition

In this section, two pattern-recognition tasks were computer simulated to demonstrate the feasibility of using an ANN for our proposed biochip module architecture. A novel sigmoid-logarithmic function is also integrated within the learning algorithm (i.e., back-propagation algorithm) to demonstrate the capability of recognizing relatively dim patterns. The study in this section will assist our future circuit design and may contribute to the new techniques for medical image processing.

In most of the fluorescence spectroscopy applications, the fluorescence patterns usually have relatively low intensities and are difficult to analyze. We know that high-excitation intensities and long exposure time can lead to stronger fluorescence signals. However, low-energy consumption and fast detection are the design goals for our biochip module architecture. Therefore, if the posterior ANN of our biochip architecture can analyze dim fluorescence patterns better, we can potentially use relatively lower energy and shorter time to conduct the analysis task.

Regarding the neural network learning algorithm, the simplest transfer function that we can use in the algorithm is a linear ramp function (e.g., linear slope between 1 and −1, flat and continuous outside [1, −1]). However, higher recognition capability can be achieved by using nonlinear transfer function in the neural network learning algorithm.

The nonlinear sigmoid (logistic) transfer function is usually adopted in artificial neural network models because its derivative can be easily obtained algebraically. For example, we define $A(h)$ as a sigmoid function:

$$A(h) = \frac{1}{1+\exp(-h)}. \tag{1}$$

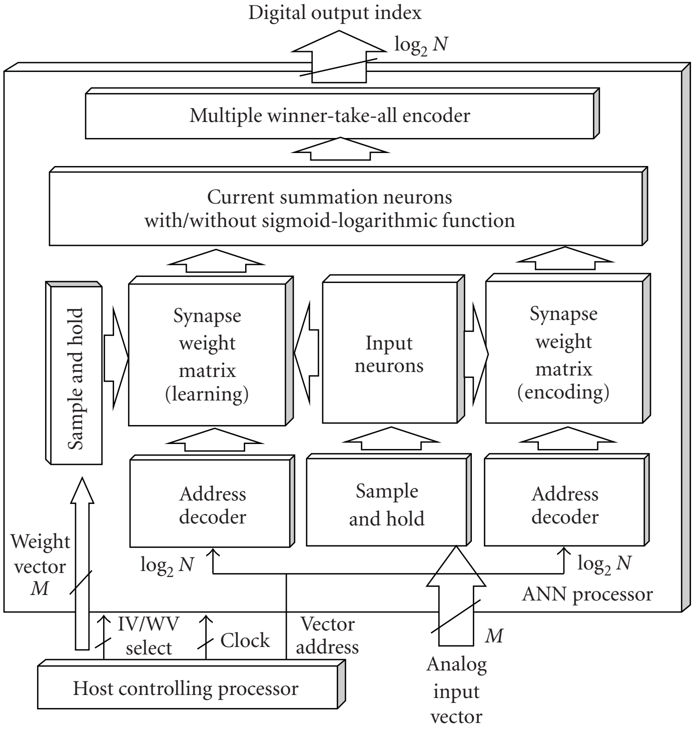


Figure 5: A system-on-chip architecture design for the winner-take-all selforganization artificial neural network chip.

Table 1: Biosignature and OCR recognition results (Unit: counts of patterns correctly recognized in one test dataset. OCR: each test dataset contains 100 characters. BIO: biosignature recognition task, each test dataset contains 20 patterns.)

| Gray level to original data | Transfer function (learning rate)<br>Hybrid sigmoid-logarithmic<br>($\eta = 0.03$) | | $1/(1 + \exp(-h))$<br>($\eta = 0.03$) | | $1/(1 + \exp(-5h))$<br>($\eta = 0.0018$) | |
|---|---|---|---|---|---|---|
| | BIO | OCR | BIO | OCR | BIO | OCR |
| 1 (original) | 20 | 64 | 20 | 79 | 20 | 63 |
| 1/3.16 | 20 | 58 | 6 | 48 | 13 | 65 |
| 1/10 | 20 | 58 | 1 | 25 | 1 | 37 |
| 1/31.6 | 20 | 47 | 1 | 25 | 1 | 25 |
| 1/100 | 19 | 32 | — | — | — | 25 |
| 1/316 | 17 | 26 | — | — | — | — |
| 1/1000 | 17 | 25 | — | — | — | — |
| 1/10000 | 17 | 25 | — | — | — | — |

The first derivative of $A(h)$ can easily be calculated by using the identity $A'(h) = A(A - 1)$. Therefore, computation complexity and cost of hardware or software can be reduced.

In addition, a single-layer feedforward network (SLFN) with any bounded continuous nonconstant activation (transfer) function or arbitrarily bounded activation (transfer) function with unequal limits at the infinities can form decision regions with arbitrary shapes [16–18]. A multilayer perceptron architecture naturally consists an SLFN in its structure. As long as a function has unequal upper bond, lower bond, and monotonic behavior, it can be used as a transfer function.

For the above computational advantage and theoretical reasons, we proposed a novel piecewise sigmoid-logarithmic function that also yields similar mathematical identities and computational benefits:

$$A(h) = \begin{cases} \dfrac{1}{1 + \exp(-h)}, & h < -2, \\ -\alpha \ln(\beta(\delta - h)), & -2 \le h < 0, \\ \alpha \ln(\beta(h + \delta)), & 0 \le h < 2, \\ \dfrac{1}{1 + \exp(-h)}, & 2 \le h. \end{cases} \quad (2)$$

In this piecewise function, $\alpha = 0.050095635$, $\beta = 1000$, $\delta = 0.01$, and $h$ is the net weighted input to the transfer function $A(h)$. The central part (net input ranging from $-2$ to 2) of the original sigmoid function was replaced by two asymmetric pieces of logarithmic curves.

To demonstrate the capability of recognizing dim patterns by using a feedforward ANN with sigmoid-logarithmic transfer function, a simple pseudo genetic assay analysis task and an optical character pattern-recognition task were simulated. MATLAB programs were created to train an ANN and examine its performance.

A 100-100-2 (100 inputs, 100 hidden neurons, and 2 output neurons) artificial feedforward neural network was chosen to perform both recognition tasks. For the biosignature recognition, 20 patterns/clusters on a microchip genetic assay were prepared (Figure 6). For simplicity, we used the pixellated image of a fluorescence image of the sample material to represent a normalized (i.e., after taking logarithmic value of the ratio of the sample to reference signals) but noisy image for the analog input to the ANN. Additional seven datasets by rescaling the gray level of the original dataset with different rescaling factors were also prepared (factors: 1/3.16, 1/10, 1/31.6, 1/100, 1/316, 1/1000, 1/10000 of the original gray level). Noticing that, both brightness and contrast levels of these new biopatterns were reduced by the rescaling factor. The three desired biopatterns (solid framed in Figure 6 that were randomly picked) are what we were searching in a scenario of finding the designed PCR assay cluster pattern of the subject with certain disease.

Similarly, in the optical character-recognition task, a dataset containing 100 different alphanumeric letters 1, 2, 3, and 4 was prepared first (as shown in Figure 7). By using the same rescaling factors (as listed in the biopattern recognition), we generated the other seven datasets that contain dimmer hand script patterns.

In both experiments, the digitized biopattern and character datasets were used for both training and testing the artificial neural network. In contrast to the traditional method of preparing independent training and test datasets, the test datasets were assigned to be identical to the training datasets in order to examine the feasibility of the proposed ANN model with the sigmoid-logarithmic function.

For simplicity, the intensities of the high-resolution pixels of each original fluorescence dot in Figure 6(a) were averaged and replaced by a single super pixel in the pixellated image in Figure 6(a). If the patterns with the lowest resolution (one pixel for one dot/well) can be correctly recognized by the ANN, the patterns with higher resolution should be recognized by the ANN with higher recognition accuracy. To economically implement the bioimaging chip, one neuron unit would be sufficient for receiving all the lights emitting from the fluorescence profile of the imaging dot/well. The results of this simulated ANN model would assist the future design of the proposed architecture. The average of the intensities of the pixels of the original fluorescence patterns is considered simple yet physically reasonable.

The back-propagation training using sigmoid-logarithmic transfer function and gradient descent method was conducted to find a convergent weight configuration (with fixed learning rate $\eta = 0.03$). A criterion is employed to count the percentage of training data that has been learned with an error less than 20%. In order to guide the weight configuration closer to a convergence condition, the BP training using regular sigmoid function was conducted first. The BP training using sigmoid-logarithmic function was conducted afterwards.

The entire procedure of the BP training algorithm using sigmoid-logarithmic transfer function is described as follows.

(i) Prepare the input patterns for the feedforward multilayer perceptron (MLP) neural network.

(ii) Assign the target values for the associated input patterns.

(iii) Use the input patterns to train the multilayer perceptron with sigmoid transfer function until the criterion value becomes close to one. Now the weight configuration is closer to a convergence condition for latter training.

(iv) Use the weight values obtained in the previous step as the initial weight condition for training the multilayer perceptron with the logarithmic-sigmoid transfer function. The regular BP algorithm using the gradient descent method is again adopted. After the criterion becomes one, the training is finished.

(v) Use this trained MLP with logarithmic-sigmoid transfer function to recognize the test data set. Examine the recognition accuracy.

The detailed conditions and pseudocodes of the BP algorithm with sigmoid and logarithmic-sigmoid transfer function are provided as the following.

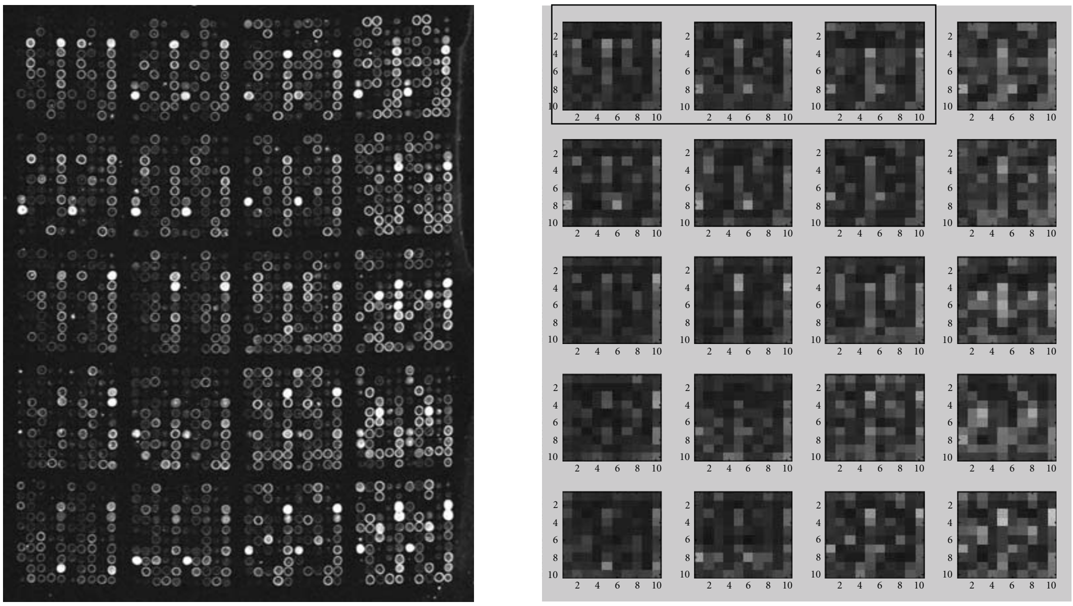
The initial weight values for the first weight matrix $W$ and second weight matrix $V$ were given randomly. The range of these random weight values was set between $-1$ and 1. However, the range of the trained weight values was unlimited. The input vector $X$ was the vectorized pixellated pattern. The associated target values $D$ were given. The learning rate (coefficient in front of the gradient partial derivative) is parameter $\eta$. The number of iterations is parameter *iter*. The training data set was used for testing as well to verify the feasibility of this proposed algorithm.

The pseudocode for the regular back-propagation algorithm using sigmoid transfer function is listed in Algorithm 1.

The pseudocode for the back-propagation algorithm using the piecewise sigmoid-logarithmic transfer function is listed in Algorithm 2.

### 4.2. Simulations results

The result of recognizing all of the normalized genetic assay datasets by the trained 100-input-100-hidden-neuron-2-output-neuron network is shown in Table 1 (BIO). The network using new transfer function can still find one desired biosignature in the dataset with factor 1/100 of the original gray level. However, the network using conventional sigmoid



(a) (b)

Figure 6: (a) Fluorescence image of a sampled microarray of cDNA, Cy3 dye, and Cy5 dye mix (only Cy5 red fluorescence is shown) [19]. Each cluster consists of a 10 × 10 grid of sample dots. Each dot corresponds to the location of a cDNA probe to which mRNA from the cells of interest have been bound. (b) Pixellated images of the clusters from the left microarray photo. Each cluster consists of a 10 × 10 pixel array that mimics a normalized biosignature pattern. The first three randomly picked desired patterns (enclosed in the solid line frame) have indices (0, 0), (0, 1), and (1, 0), accordingly. The rest unwanted patterns share index (1, 1).

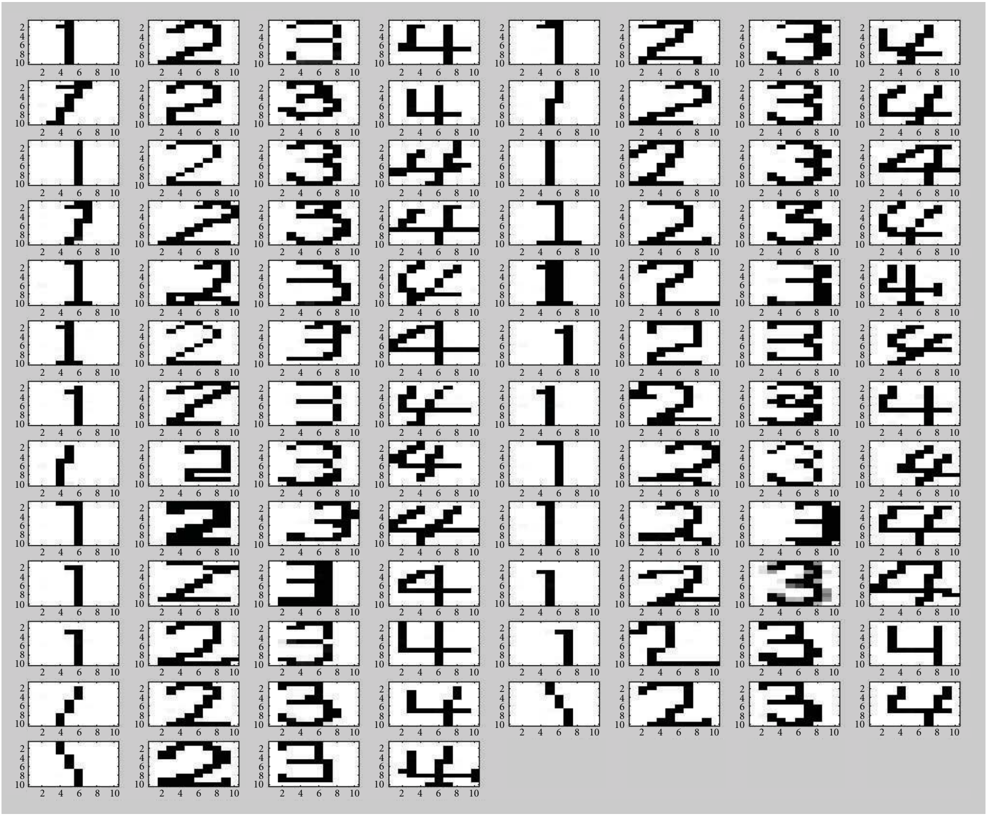


Figure 7: The picture of 100 different patterns of alphanumeric letters 1, 2, 3, and 4 used in the optical character recognition experiment.

```
Define input vector to the first layer that contains input
patterns $X$ with $-1$ bias
For $i = 1$ to $iter$
    Calculate hidden neuron output using
    sigmoid function
    Define output vector from the hidden layer
        that contains hidden neuron output and $-1$ bias
    Calculate the final output using sigmoid function
    Check the criterion of percentage of the input data
                    that has an error less than 20%
    If all input data have errors less than 20%, stop
    the training
    Compute the first back-propagation error set
    Compute the second back-propagation error set
    Update the second weight matrix
    Update the first weight matrix
End
```

Algorithm 1

```
Define input vector that contains input patterns $X$
with $-1$ bias
For $i = 1$ to $iter$
    Calculate the hidden neuron output according to where
    the net weighted input is falling in the range of the
    piecewise sigmoid-logarithmic function
    Define output vector from the hidden layer that
    contains hidden neuron output and $-1$ bias
    Calculate the final neuron output, the first back-
    propagation error set, and the second back-propagation
    error set according to where the net weighted input is
    falling in the range of the piecewise sigmoid-logarithmic
    function
    Check the criterion of percentage of the input data
                    that has an error less than 20%
    If all input data have errors less than 20%,
    stop the training
    Update the second weight matrix
    Update the first weight matrix
End
```

Algorithm 2

transfer functions cannot distinguish biopatterns well in the datasets with factor 1/3.16 of the original gray level or below. Consistent recognition accuracy was obtained for the optical character recognition (OCR) tasks (Table 1 (OCR)). The network using new transfer function can recognize 32 characters in the dataset with factor 1/100 of the original gray level but not the network with regular sigmoid function. The recognition accuracies for datasets with a factor below 1/316 become constant because the network tends to recognize one category perfectly according to the converged weight configuration in this particular experiment.

## 5. CONCLUSION

A new optoelectronic multichip microsystem for real-time field applicable robust dual-label PCR assay analysis was proposed. This microsystem architecture contains a front-end bioimage chip for analog signal conversion and augmentation, and an artificial neural network for the autonomous data analysis purpose. A differential logarithmic bioimage chip is designed and presented. The typical data analysis procedure of taking logarithm of the ratio of the normalized post-PCR sample intensity is conducted effectively in this differential logarithmic bio-image chip. A single channel logarithmic circuit of the differential logarithmic bioimage chip was designed, fabricated, and characterized. The weak fluorescence signals can be amplified by this logarithmic amplifier circuit for easier data analysis. Regarding the ANN subsystem, an unsupervised hardware version: a weight-reconfigurable winner-take-all ANN SoC chip suitable for selforganized Kohonen filter algorithm, and a supervised software version: a computer-simulated ANN using back-propagation algorithm with a novel sigmoid-logarithmic transfer function is presented. The back-propagation neural network learning algorithm using the sigmoid-logarithmic function was successfully simulated. The simulation results show that a trained ANN using this new transfer function can classify low-fluorescence patterns better than using the conventional sigmoid transfer function. This software model might be applicable to other medical image processing tasks. In summary, by integrating the optical setup, the bioimage chip, and the artificial neural network processor with excellent performances and advantages listed previously, we can envision the success of using this compact microsystem to conduct on-site, real-time, noise-tolerable, and high-throughput dual-labeled genetic expression analysis efficiently.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Armand R. Tanguay, Jr. at University of Southern California for the kind instructions and funding for implementing the logarithmic amplifier circuitry.

## REFERENCES

[1] E. T. Lagelly, J. R. Scherer, R. G. Blazej, et al., "Integrated portable genetic analysis microsystem for pathogen/infectious disease detection," *Analytical Chemistry*, vol. 76, no. 11, pp. 3162–3170, 2004.

[2] N. Agrawal, Y. A. Hassan, and V. M. Ugaz, "A pocket-sized convective PCR thermocycler," *Angewandte Chemie International Edition*, vol. 46, no. 23, pp. 4316–4319, 2007.

[3] T. M.-H. Lee, M. C. Carles, and I.-M. Hsing, "Microfabricated PCR-electrochemical device for simultaneous DNA amplification and detection," *Lab on a Chip*, vol. 3, no. 2, pp. 100–105, 2003.

[4] M. W. Craven and J. W. Shavlik, “Machine learning approaches to gene recognition,” *IEEE Expert*, vol. 9, no. 2, pp. 2–10, 1994.

[5] X. Guan, R. J. Mural, J. R. Einstein, R. C. Mann, and E. C. Uberbacher, “GRAIL: an integrated artificial intelligence system for gene recognition and interpretation,” in *Proceedings of the 8th Conference on Artificial Intelligence for Applications*, pp. 9–13, Monterey, Calif, USA, March 1992.

[6] K. Dobbin, J. H. Shih, and R. Simon, “Questions and answers on design of dual-label microarrays for identifying differentially expressed genes,” *Journal of the National Cancer Institute*, vol. 95, no. 18, pp. 1362–1369, 2003.

[7] T. Kohonen, *Self-Organization and Associative Memory*, Springer, New York, NY, USA, 3rd edition, 1989.

[8] P. J. Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, John Wiley & Sons, New York, NY, USA, 1994.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[10] A. Zöller, R. Götzelmann, K. Matl, and D. Gushing, “Temperature-stable bandpass filters deposited with plasma ion-assisted deposition,” *Applied Optics*, vol. 35, no. 28, pp. 5609–5612, 1996.

[11] S. G. Chamberlain and J. P. Y. Lee, “A novel wide dynamic range silicon photodetector and linear imaging array,” *IEEE Journal of Solid-State Circuits*, vol. 19, no. 1, pp. 41–48, 1984.

[12] C. A. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, Mass, USA, 1989.

[13] J.-C. Lue, “Neuron unit arrays and nature/nurture adaptation for photonic multichip modules,” Ph.D. dissertation, University of Southern California, Los Angeles, Calif, USA, 2007.

[14] W.-C. Fang, B. J. Sheu, O. T.-C. Chen, and J. Choi, “A VLSI neural processor for image data compression using self-organization networks,” *IEEE Transactions on Neural Networks*, vol. 3, no. 3, pp. 506–518, 1992.

[15] W.-C. Fang, “A smart vision system-on-a-chip design based on programmable neural processor integrated with active pixel sensor,” in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS ’00)*, vol. 2, pp. 128–131, Geneva, Switzerland, May 2000.

[16] G.-B. Huang, Y.-Q. Chen, and H. A. Babri, “Classification ability of single hidden layer feedforward neural networks,” *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 799–801, 2000.

[17] I. W. Sandberg, “General structures for classification,” *IEEE Transactions on Circuits and Systems I*, vol. 41, no. 5, pp. 372–376, 1994.

[18] K. M. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[19] L. Hu, J. Wang, K. Baggerly, et al., “Obtaining reliable information from minute amounts of RNA using cDNA microarrays,” *BMC Genomics*, vol. 3, no. 16, pp. 1–8, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 526343, 11 pages
doi:10.1155/2008/526343

*Research Article*

# Textural Classification of Mammographic Parenchymal Patterns with the SONNET Selforganizing Neural Network

**Daniel Howard,[1] Simon C. Roberts,[1] Conor Ryan,[2] and Adrian Brezulianu[3]**

[1] *QinetiQ, St Andrews Road, Malvern, Worcestershire WR14 3PS, UK*
[2] *Department of Computer Science and Information Systems, College of Informatics and Electronics, University of Limerick, Ireland*
[3] *Faculty of Electronics and Telecommunications, "Gh.Asach" Technical University of Iasi, 700050 Iasi IS, Romania*

Correspondence should be addressed to Daniel Howard, dr.daniel.howard@gmail.com

Received 12 September 2007; Accepted 16 January 2008

Recommended by Halima Bensmail

In nationwide mammography screening, thousands of mammography examinations must be processed. Each consists of two standard views of each breast, and each mammogram must be visually examined by an experienced radiologist to assess it for any anomalies. The ability to detect an anomaly in mammographic texture is important to successful outcomes in mammography screening and, in this study, a large number of mammograms were digitized with a highly accurate scanner; and textural features were derived from the mammograms as input data to a SONNET selforganizing neural network. The paper discusses how SONNET was used to produce a taxonomic organization of the mammography archive in an unsupervised manner. This process is subject to certain choices of SONNET parameters, in these numerical experiments using the craniocaudal view, and typically produced O(10), for example, 39 mammogram classes, by analysis of features from O($10^3$) mammogram images. The mammogram taxonomy captured typical subtleties to discriminate mammograms, and it is submitted that this may be exploited to aid the detection of mammographic anomalies, for example, by acting as a preprocessing stage to simplify the task for a computational detection scheme, or by ordering mammography examinations by mammogram taxonomic class prior to screening in order to encourage more successful visual examination during screening. The resulting taxonomy may help train screening radiologists and conceivably help to settle legal cases concerning a mammography screening examination because the taxonomy can reveal the frequency of mammographic patterns in a population.

## 1. INTRODUCTION

Nationwide mammography screening (NMS) is the most successful method for the early detection of breast cancer [1]. There exist differing opinions concerning the details of an ideal NMS programme, with a body of opinion arguing that it should involve women over the age of 40 attending an annual mammography examination that obtains X-ray of each breast (for the two standard views), with technicians ensuring that the subject is imaged consistently each year to allow a comparison of the images over time, which is essential to mammography screening. Longitudinal comparisons of mammograms are quite revealing. Consider Figure 1, which pertains to a woman who assisted frequent screening for over 20 years. This longitudinal comparison llustrates the involution of the parenchyma with age of the subject as it is replaced by adipose tissue. Exceptions relate to extensive fibrosis or adenosis for which this involution to all intent and purposes cannot be observed [2]. Also, mammograms of subjects starting hormone replacement therapy appeared to us to restore an earlier appearance. By comparing images longitudinally, it may be possible to detect a lesion as an abnormality that grows in time in contrast to the gradual but perceivable retraction of the parenchyma.

An NMS programme also requires an integrated clinical team involving a pathologist, radiologist, oncologist and surgeon. Ideally, when a patient undergoes an invasive procedure then the radiologist should obtain X-ray images of sections of the biopsy specimen to visually compare these with the original mammogram so as to improve his or her abilities at detecting lesions.

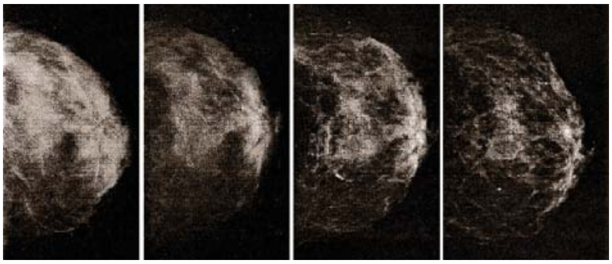


Figure 1: Longitudinal study showing the process of involution of the parenchyma with age of subject (digitized from L. Tabár archive). The CC view is shown from examinations in 1980, 1989, 1997, and 2002 (left to right).

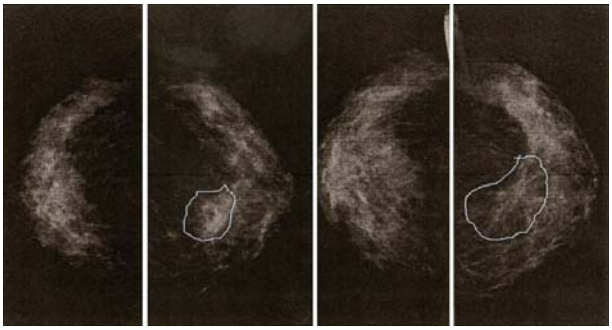


Figure 2: Left to right: longitudinal progression of a distortion (digitized from L. Tabár archive).

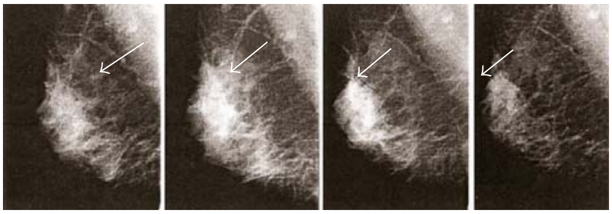


Figure 3: Left to right: longitudinal progression of a distortion (digitized from L. Tabár archive).

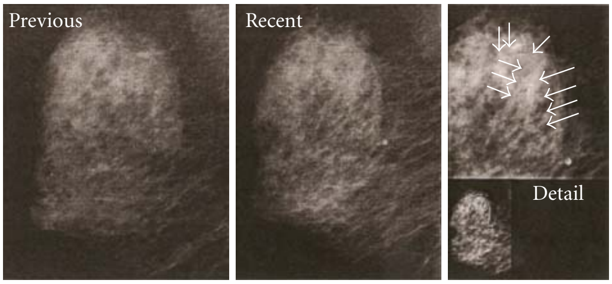


Figure 4: Left to right: longitudinal progression of a distortion (digitized from L. Tabár archive).

If the NMS programme were to encourage the same radiologist to screen the same women for a number of years, it is possible that fewer women would be recalled without good reason. When fewer women are recalled without good reason, higher levels of compliance and participation of women usually follow, and if so, then the benefits of screening: early detection and arguably lower cancer mortality rates should follow.

To date, computer-aided detection (CAD) of cancer does not appear to have achieved significant penetration in NMS services. A challenge of CAD is to help with difficult-to-perceive and subtle lesions also known as distortions, which are strong indications of breast cancer. Calcifications are not as significant because up to 80% of calcification occurrences in screening are benign [3].

Architectural distortions are variable in shape and size and difficult to pick-up in early detection even by an experienced radiologist. Figures 2, 3, and 4 illustrate this typical progression for consecutive examinations over time, the X-ray images can be seen to differ most significantly with an impression of a "pulling" effect at the site of the distortion. These examples are very obvious but it is necessary to detect very subtle distortions early on. By observing the foremost radiologists at work it became apparent that anomalies in image texture (angles in the orientation of textural patterns) accounted for their intuitions about the presence of a lesion. Radiologists have used emotive terms to explain how they pick-up on subtle anomalies and the detection of mammographic lesions has been described as "visual art" [3]. This together with the variability and elusiveness of architectural distortions motivated us to develop machine vision algorithms to construct a taxonomy of mammography textural patterns rather than to produce a more standard CAD tool. It was observed that unusual angular orientations of texture alerted the experienced radiologist to a subtle distortion. This observation, however, is an oversimplification for the purpose of this discussion, as clinical knowledge is essential for the detection of the lesion.

## 2. A COMPUTATIONAL TOOL FOR MAMMOGRAM CLASSIFICATION

How to best assist an NMS programme using a computational tool for image analysis? This is the research question that we set out to answer. In cancer screening, the number of normal cases far exceeds the number of suspected lesions (perhaps to a ratio of 50 to 1). This can lead to fatigue in human-facilitated screening and a lack of lesion examples with which radiologists are trained. Furthermore, the high variability of normal cases has led it to be said that a mammogram can equate to a fingerprint in its subject identification ability.

Our aim was to apply artificial intelligence (AI) and image analysis techniques to answer the aforementioned research question. In order to study the problem quantitatively, we deployed an accurate Lumysis scanner to digitize a large number of pristine film mammograms from a long-established archive that contains arguably the greatest density of high-quality mammograms in the world [4]. Leading radiologists of the breast such as Wolf and Tabár considered "parenchymal types" [3, 5] to establish a subjective classification of mammograms into a few categories. This paper addresses the automation of this classification process by using an unsupervised, selforganizing method of AI known as SONNET (described in Section 3). The automated classification scheme is based on image textural analysis because, as previously mentioned, we surmised that textural patterns and orientations (the angles that can be perceived in this texture) are the important criteria of visual inspection for radiologists; they constitute a type of visual "algedonic alert" [6] to the presence of a subtle distortion in the mammogram.

A categorized organization of a mammogram archive by texture could have many potential uses. Learning and training are obvious but for example with the objective of higher accuracy in screening, the mammograms could be sorted by taxonomic pattern to enable the organized viewing that gives time for radiologists to become accustomed to the background texture. This normal texture could then more easily be compared against the anomalous texture of architectural distortions and lesions. Also, patients could be tracked longitudinally based on their progression through the taxonomy of parenchymal types [4, 5, 7]. This could provide additional information for a mammographic examination by allowing comparison to other patients who experienced a similar progression. The aberrations of normal breast development may be studied with this taxonomy. They could offer clues to reasons for a higher incidence of breast cancer in a population, for instance, according to lifestyle or genetic profile. Conceivably, a taxonomy could help to settle a legal case as it quantifies how common is the parenchymal pattern, to help settle a dispute concerning an early breast cancer warning that was missed.

## 3. THE SONNET CLASSIFIER

The artificial neural network known as SONNET [8] consists of an array of classifiers connected to an input field as shown in Figure 5. Input patterns are presented in turn to the input field and *typical* patterns are gradually encoded as follows. The constituent classifiers compete to encode each pattern such that the classifier with the best match to the current input tends to adapt itself more than the other classifiers. This *winning* classifier adapts itself by partially encoding the current input pattern on *weighted* excitatory connections from the input field. Furthermore, the classifier adapts weighted inhibitory connections to the other competing classifiers, thus allowing the winning classifier to suppress its competitors. The classifiers consequently diverge so that each responds only to input patterns which are similar to the pattern encoded on its excitatory connections.

SONNET is a selforganizing neural network based on adaptive resonance theory [2] that encodes classifications using unsupervised learning. The SONNET architecture is shown in Figure 5 where a field of input neurons is connected to a field of classification neurons via weighted connections. An input pattern is a pattern of relative neural activity in the input field at any given moment, and a set of input patterns is presented to SONNET by setting the activity of the input neurons to each pattern in turn for a fixed duration.

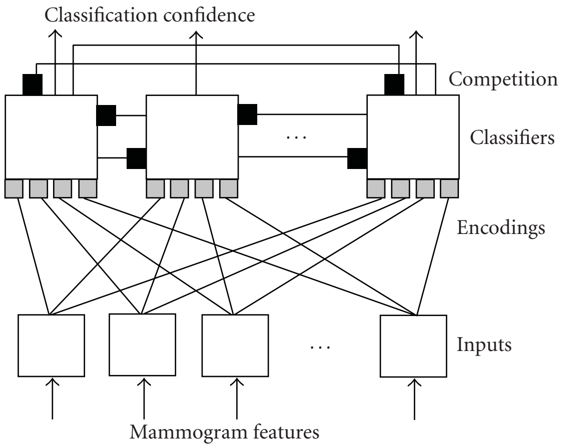


Figure 5: SONNET architecture.

Each input neuron is connected to each classification neuron by an excitatory weighted connection, and the weight on each connection is continually adapted via a learning rule (this is similar to Hebbian learning [9, 10] which postulates that memory is stored in the synaptic weights and learning is the process that changes those weights) such that the connection becomes stronger when the two corresponding neurons are simultaneously active. Furthermore, higher activity on the two neurons causes the connection to become stronger more rapidly. The maximum rate at which the weights can change governs the network's learning speed and this is regulated by controlling parameters that are set prior to a SONNET run.

The relative pattern of excitatory weights on connections to a single-classification neuron represents the so-called *prototype* for that classifier. The excitatory input to a classification neuron is based on two measures. The first measure is based on the size of the excitatory weights so that a large excitatory input can be achieved when strong weights gate high-input activations. The second measure quantifies how well the prototype matches the current input pattern such that a large excitatory input can be achieved for a good match even when the prototype is represented by small excitatory weights. A large excitatory input to a classification neuron allows the neuron to gain a high activation in response to the current input pattern. This activation represents the confidence with which the neuron classifies the input pattern. The learning speed can be set to allow the prototype to form gradually from repeated exposure to input patterns, such that the prototype encodes a *generalization* for multiple similar input patterns. The classification neuron can then obtain a high activation when any one of these input patterns occurs and thus it classifies these patterns together.

The classifiers compete to encode each input pattern such that the classifier with the best match to the current input tends to adapt itself more than the other classifiers, thus further improving its competitive advantage. Each classifier

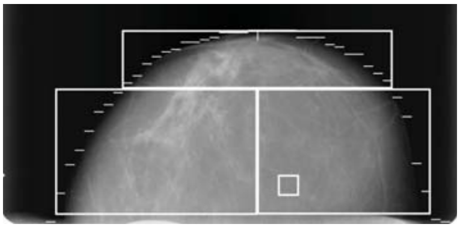


Figure 6: Mammogram regions.

is connected to all other classifiers by an inhibitory weighted connection that is again adapted via a learning rule such that a connection becomes stronger (i.e., more inhibitory) when the corresponding neurons are simultaneously active. Classifiers that have partially encoded similar patterns thus compete strongly against each other, causing one classifier to eventually suppress its competitors. The classifiers consequently diverge to encode different input patterns so that only one classification neuron achieves a high activation in response to each input pattern. When the input pattern changes, the activation of a previously excited classifier can decrease due to both passive decay and inhibition from competing classifiers that better represent the new input pattern.

SONNET performs real-time learning by continually adapting its weights in the selforganizing manner described above. The learning algorithm is unsupervised and there is no reinforcement of any kind from an external source to judge the emergent classifications against expected classifications. The network is initialized with random small weights and the classifiers compete such that the most common input patterns are encoded first and the less common patterns are encoded more gradually.

The selforganizing behaviour causes SONNET to be susceptible to the so-called *stability-plasticity dilemma* [2], which states that a network should always remain adaptive to learn new patterns (i.e., have plasticity) without degrading well-formed encodings for previously learned patterns (i.e., have stability). SONNET achieves plasticity due to the aforementioned learning algorithm but it also achieves stability by reducing the learning speed at a single classifier when the size of the classifier's excitatory weights become large. A classifier can only gain large excitatory weights after it has encoded a good representation for one or more input patterns, and a stable classifier is said to be *committed* with excitatory weights that constitute a long-term memory of the encoding.

For the current application, each input pattern represented features extracted from a mammogram. A set of mammograms was selected with which to *train* SONNET, and each presentation of the full mammogram set is known as an epoch. SONNET typically learned by adapting itself over many epochs until a *stable* set of classifiers could classify each mammogram with a significant degree of confidence. The order of mammogram presentation was randomized on each epoch. This reduced the likelihood of an unstable classifier from oscillating between similar yet significantly different potential classes.

SONNET is a highly dynamic system which is controlled by many parameters as discussed in other recent research presentations in [11–13]. It is a fully unsupervised system which encodes classes via selforganization in response to the input patterns. However, the manual specification of SONNET's controlling parameters allows a degree of supervision. For example, a number of parameters govern SONNET's learning speed which in turn influences the number of classes encoded. The greatest learning speed produces *one-shot learning* where SONNET simply memorizes each input pattern. Slower learning produces broader classes, where a single classifier can represent multiple similar patterns by forming encodings that generalize the characteristic features of the class.

Multiple SONNET runs were conducted using different randomized initial weights on the connections within the network. This allowed different encodings to form on each run. SONNET's controlling parameters were also varied on different runs to change the learning speed. SONNET comprised at most 80 classifiers though the actual number was set in accordance with the learning speed. Each run terminated after 100 epochs but the final epoch did not necessarily represent the optimum SONNET state. Section 4.5 explains how the optimum SONNET runs and epochs were identified.

## 4. DEVELOPING A MAMMOGRAM TAXONOMY USING UNSUPERVISED CLASSIFICATION

The development of an unsupervised classification scheme to produce a mammogram taxonomy had to address the following issues: input feature extraction; input feature selection to produce a minimal set of features which best characterize the input cases; input feature preprocessing prior to presentation to the classification system; classifier development; and the definition of classification performance measures in order to compare the classifiers resulting from different SONNET runs. These issues are discussed in the following subsections.

### *4.1. Mammogram feature extraction*

450 mammograms were chosen for the current study. These mammograms represented the CC left and right views for 225 different patients. The mammograms were X-rayed between 1990 and 2002; and they were of a highly consistent top quality. Most mammograms displayed normal breast tissue but 49 of the patients had been diagnosed as having breast cancer. Subtle cancerous lesions were evident in the mammograms corresponding to these patients.

The breast tissue in a mammogram must be segmented from the background before mammogram features can be extracted. This was achieved by locating maximal brightness gradients to produce multiple hypotheses for the actual breast margin. The best hypothesis was identified by optimizing contour shape and smoothness. The location of the nipple was also estimated to ascertain three different regions within the breast as shown in Figure 6. These regions are the retroareolar region (behind the nipple), an axillar region

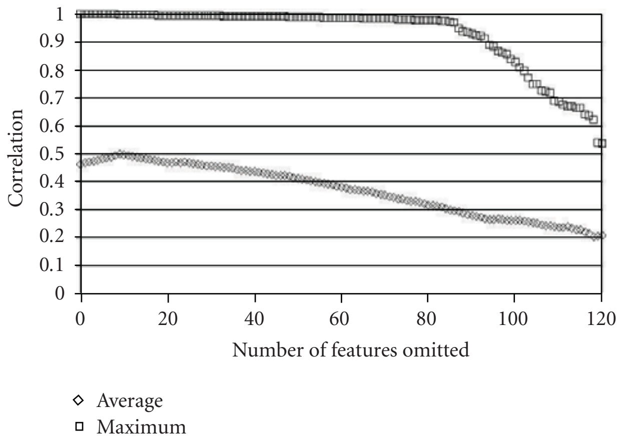


FIGURE 7: Average and maximum correlations between mammogram features against the number of features omitted.

(outer) and a medial region (inner). The identification of the breast margin allowed equivalent regions to be defined on different mammograms by specifying positions relative to the nipple location.

Standard image processing techniques were used to extract the following information from each of the three regions: brightness distribution, contrast distribution, and textural measures. Brightness was calculated in a 10 mm square (as depicted in Figure 6) which was swept over each region to give the brightness distribution as represented by minimum, maximum, average, and standard deviation values. The same procedure was used to calculate the contrast distribution.

Textural measures were calculated by accumulating co-occurrence matrices over each region. The following 9 textural features were calculated from each matrix: angular second moment, inverse difference moment, contrast, entropy, sum entropy, difference entropy, and three correlation measures. Furthermore, co-occurrence matrices were generated for each of four orientations: 0, 45, 90, and 135 degrees. Hence 12 matrices were generated; four orientations for the three regions.

The above processing resulted in 132 image features for each mammogram. Note that the mammogram for the left breast was flipped horizontally before processing to map the axillar and medial regions onto those for the right breast, and to give the appropriate orientations for the textural features.

### 4.2. Mammogram feature preprocessing

The extracted image features constitute an input feature vector that can be presented to SONNET's input field. However, each dimension of the input feature vector must be normalized so that each feature varies over the same range. This prevents individual dimensions from dominating the input feature space. For example, suppose dimension $X$ ranged from 0 to 255 and dimension $Y$ ranged from 0 to 1, then without normalization $X$ would dominate $Y$ in the input vector so that $Y$ would effectively be negligible. Furthermore, the normalization improves the discrimination between input cases. In the above example, without normalization each input case would typically be represented by a vector where $X$ is two orders of magnitude greater than $Y$. Consequently, the input cases would appear more similar to each other than if each input dimension was normalized.

The mammogram features were linearly scaled to range from 0 to 1 by analyzing the mammogram set for each input dimension independently. For a single input dimension, the minimum and maximum values across the mammogram set were discovered and these were used to normalize each input case.

### 4.3. Mammogram feature selection

The 132 features extracted from each of the 450 mammograms were analyzed to produce a minimal set of features which best characterized the mammograms. The procedure for this was as follows:

(i) calculate the correlation between each pair of features across the set of mammograms,

(ii) identify the most correlated pair of features; features $X$ and $Y$,

(iii) omit feature $X$ if it has the least deviation across the set of mammograms, else omit feature $Y$,

(iv) repeat from step 2 whilst the highest correlation is above a prescribed threshold.

This procedure produced the correlations shown in Figure 7. It can be seen from the maximum correlations that many of the features were highly correlated. These correlations corresponded to the same type of textural features taken from the same mammogram regions, but where the features pertained to different textural orientations. For example, the entropies in the retroareolar region at 45 degrees and 135 degrees were highly correlated. The maximum correlation between nontextural features was 0.87.

The figure shows that the omission of highly correlated features tended to reduce the average correlation between features after an initial increase in this average. The discrimination between mammograms improves as the average correlation between the features is minimized. However, as the average correlation tends to continually decrease a correlation threshold must be set to terminate feature omission. This threshold was set by considering the distance between mammograms in feature space.

Section 4.1 explained that each feature was scaled to range from 0 to 1, hence the maximum distance between two mammograms in feature space was the square root of the number of features used. For example, the maximum distance for the original 132 features was 11.5. Therefore, for a given number of features, the distance between mammograms can be calculated and then normalized by the maximum potential distance.

Figure 8 displays the variation in the average normalized distance between mammograms as highly correlated features were omitted. Similarly, Figure 9 displays the variation in

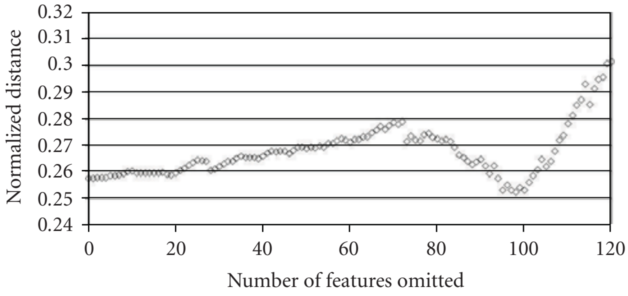


Figure 8: Average normalized distance between mammograms against the number of features omitted.

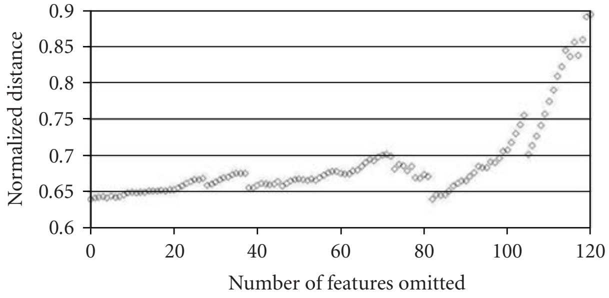


Figure 9: Maximum normalized distance between mammograms against the number of features omitted.

the maximum normalized distance between mammograms. The normalized distance between mammograms should be maximized to improve the discrimination between mammograms. The figures show that the normalized distances increase slightly as more features were omitted. However, excessively omitting features would restrict the information captured from the mammograms. Consequently, a correlation threshold of 0.98 was set to terminate feature omission. This caused 79 features to be omitted which approximately corresponds to local maxima in Figures 8 and 9.

In summary, mammogram features were omitted from the input vector in order to minimize the correlation between the remaining features, whilst maximizing the normalized distance between the mammograms in feature space. A correlation threshold of 0.98 was set to limit the highest correlation between mammogram features. This caused 79 features to be omitted and thus retained 53 features. SONNET's input field therefore consisted of 53 input neurons where each neuron represented a specific type of mammogram feature.

### 4.4. Classification performance measures

This section defines performance measures to compare different mammogram classifications. The measures in the first subsection are general to any classification task whereas those in the second subsection are specific to mammogram classification.

#### 4.4.1. Distance in input feature space

A set of input cases can be conceived as a set of points in input feature space. Thus the performance of a classification scheme can be quantified by considering the distances between input cases in input feature space. These distances give rise to the following rule. *Input cases which receive the same classification should be proximate in feature space, whereas cases which are classified differently should be distant from each other.* Hence, the classification task becomes a multiobjective optimization problem which is required to minimize the average within-class distance between case-pairs, whilst maximizing the average between-class distance.

Performance measures can be formulated for the current task by considering two mammograms $i$ and $j$ which are a distance $d_{ij}$ apart in input feature space. Suppose that these mammograms are classified as being of type $\chi_i$ and $\chi_j$ respectively, and that the corresponding classification confidences are $c_i$ and $c_j$. It is more important for mammograms which are classified with a high confidence to be consistent with the above rule, than it is for mammograms classified with a lower confidence. Hence, the distance $d_{ij}$ should be weighted by the confidences $c_i$ and $c_j$.

The average distance over a set of mammograms can now be calculated. The average within-class distance, $D_w$, would be calculated over the set of mammograms which received the same classification (i.e., $\chi_i = \chi_j$), whereas the average between-class distance, $D_b$, would be calculated over the set of mammograms which received different classifications (i.e., $\chi_i \neq \chi_j$). These average distances are calculated as follows:

$$D_w = \frac{\sum_{\chi_i=\chi_j} c_i c_j d_{ij}}{\sum_{\chi_i=\chi_j} c_i c_j}, \tag{1}$$

$$D_b = \frac{\sum_{\chi_i\neq\chi_j} c_i c_j d_{ij}}{\sum_{\chi_i\neq\chi_j} c_i c_j}. \tag{2}$$

#### 4.4.2. Patient-wise mammogram comparison

A patient should receive the same classification for their left and right CC mammograms, and this notion was confirmed by casual subjective observation. This notion can be tested by analyzing the distances between pairs of mammograms in feature space. Figure 10 shows the distances between mammogram-pairs in the reduced feature space of 53 feature types. Comparison between the left and right mammograms for the same patient produced the lower line, where each point corresponded to a single patient and the points were ranked according to increasing distance.

Each "diff D1" point was produced by comparing the right views between two different patients, and this comparison was repeated for all combinations of patient-pairs. Similarly, the "diff S1" points were produced by comparing pairs of left views of different patients. The points were again ranked according to increasing distance. There was no significance in comparing the left or right views individually and so the corresponding points overlap to appear as the upper line in Figure 10.

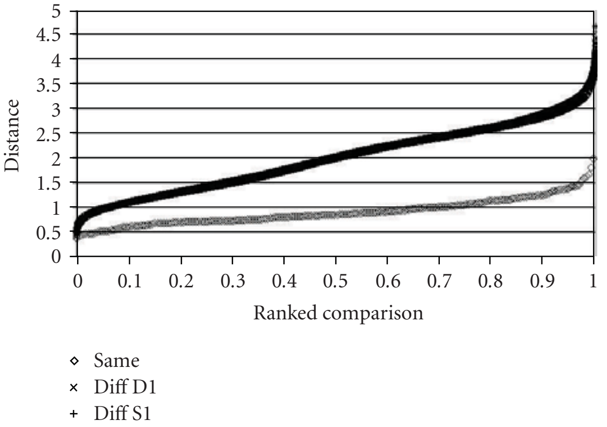


Figure 10: Comparing the distances between mammograms using 53 features: (a) *same*, the left and right views within patients, (b) diff D1, the right views between patients, and (c) diff S1, the left views between patients.

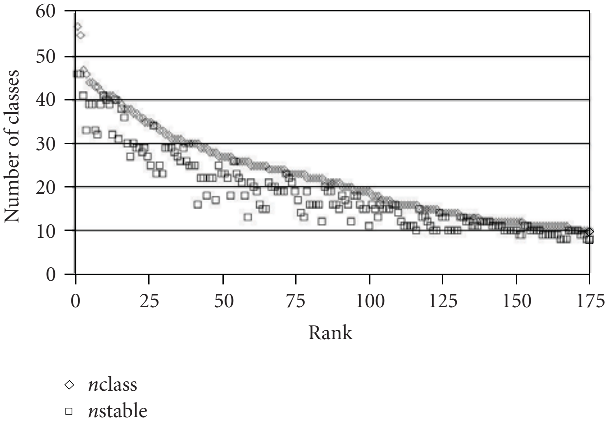


Figure 11: The number of classes formed on the best SONNET epochs. *nClass* is the total number of classes formed and *nStable* is the number of stable classes. The epochs are ranked according to descending *nClass* and then descending *nStable*.

The figure shows that within-patient distances were typically less than between-patient distances. Approximately 70% of the within-patient comparisons gave distances less than 1, and all these comparisons gave distances less than 2. Conversely, only 5% and 50% of the between-patient comparisons gave distances less than 1 and 2 respectively.

The results in this section justify the use of patient-wise performance measures for the mammogram classifiers. However, because the two mammograms for a particular patient can differ significantly, patient-wise performance measures should be used only as secondary measures. For example, patient-wise performance measures could be used to compare classifiers which are indistinguishable when using the measures based on distances in mammogram feature space. Note that patient-wise performance measures did not actively drive SONNET's development, but instead the measures were used to assess classification performance after development.

### *4.5. Discovering optimum classifications*

Section 3 stated that multiple SONNET runs were conducted for 100 epochs. Any of these epochs could represent the optimum SONNET state, where many stable classifiers separate the mammogram set into clearly distinguishable classes. Every epoch was assessed according to various performance measures and this posed a multiobjective optimization problem. The number of candidate optimum SONNET epochs was reduced by discovering the *Pareto front* across the performance measures. Consequently, none of these candidate epochs could be dominated by another epoch on *every* performance measure. The Pareto front was discovered across the following dimensions:

(i) average within-class distance $D_w$ (1),

(ii) average between-class distance $D_b$ (2),

(iii) the number of classes encoded,

(iv) the classification confidences,

(v) the fraction of patients which received the same classification for their two mammograms.

$D_w$ was minimized whereas all of the other performance measures were maximized.

## 5. MAMMOGRAM CLASSIFICATION PERFORMANCE

This section discusses the performance of SONNET in establishing a mammogram taxonomy. The optimum classifications from multiple SONNET runs were judged using the performance measures described in Section 4.5.

### *5.1. Number of mammogram classes*

Casual observation of the mammogram set can roughly indicate the number of taxonomic classes involved but it is difficult to precisely specify the number of required classes. However, the current study focused on developing a maximal number of classes to discover the typical subtleties which discriminate mammogram classes. Various SONNET parameters control the number of classes encoded. These parameters were varied to analyze the number of classes which most commonly formed, and this number was deemed to correspond to the most natural taxonomic decomposition of the mammograms.

Figure 11 displays the number of classes formed on the best SONNET epochs. These epochs relate to many different SONNET runs but a single run could also produce multiple best epochs. The epochs were ranked according to the number of classes formed and the number of these which were *stable*. The resulting rank numbers are used to identify the best epochs in the subsequent discussion.

Classes became stable in SONNET after their encoding had been refined by sufficient past experience. Unstable classes were always present however, to enable SONNET to adapt to changes in the input patterns. Therefore, the proportion of SONNET's classes which were stable represents the maturity of the overall network and the quality of the

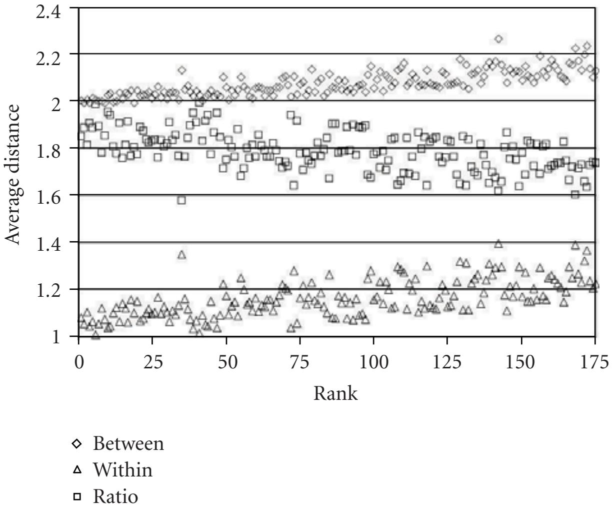


Figure 12: Average distance in input feature space for between-class mammograms and within-class mammograms for the best SONNET epochs. The ratio of between-class distance over within-class distance is also shown. The epoch ranking corresponds to Figure 11.

encodings. Hence, the best results in Figure 11 are those with the maximum number of *stable* classes.

Figure 11 shows that it was difficult for more than 40 stable classes to form. SONNET parameters were investigated to produce more classes but this resulted in SONNET memorizing individual mammograms instead of clustering them with other mammograms. SONNET commonly produced between 20 and 30 classes suggesting that this represents the most natural taxonomic breakdown. SONNET parameters could be set to form fewer, broader classes but these classifications obscure the subtleties which discriminate mammogram classes.

### 5.2. Class tightness

Figure 12 shows the average distance in input feature space for mammograms classified differently (between-class, $D_b$ given by (2)) and for mammograms classified the same (within-class, $D_w$ given by (1)). These distances were plotted for each of the best SONNET epochs which were ranked in Figure 11.

The SONNET epochs with low-rank numbers produced many narrow classes and thus yielded

(i) a low average within-class distance because mammograms had to be highly proximate in feature space to be clustered together,

(ii) a low average between-class distance because the class tightness allowed similar mammograms to be classified differently.

Conversely, SONNET epochs with high-rank numbers formed fewer, broader classes, and thus yielded higher within-class and between-class distances.

Reference distances were calculated to create a context within which to consider the between-class and within-class distances. The average distance between all the mammograms, $D_{av}$, was 2.00 and the maximum distance, $D_{\max}$, was 4.66. (These reference distances can be seen in Figure 10.) Figure 12 shows that the between-class distances for low-rank numbers approximately equalled $D_{av}$ and that the maximum between-class distance was approximately half $D_{\max}$.

A further reference distance can be calculated by considering a classification where each patient is distinct, such that their two mammograms are classified the same with a confidence of 1. This would yield $D_w = 0.90$ and $D_b \approx D_{av}$. Figure 12 shows that the within-class distances for low-rank numbers were slightly greater than 0.9, which was expected as each class clustered approximately 10 mammograms together.

The best classifications minimized the within-class distance yet maximized the between-class distance, therefore the ratio of between-class distance over within-class distance should be maximized. Figure 12 includes this ratio for each ranked epoch, and shows that the best epochs produced a ratio of almost 2 by developing relatively tight classes ($D_w \approx 1$), whilst retaining typical between-class distances ($D_b \approx D_{av}$).

### 5.3. Patient-wise performance measures

Section 4.4.2 justified the use of patient-wise performance measures to quantify the extent to which the two mammograms for each patient received the same classification. This patient-wise performance measure was used as a secondary measure to discriminate classifiers which were similar when judged using other performance measures.

Figure 13 displays the fraction of patients whose mammograms were classified the same for the best SONNET epochs. This fraction was approximately 40% for the epochs with narrow classes (low-rank numbers), as these encodings captured the subtleties which differentiate the mammograms for a single patient. Conversely, the epochs with broad classes (high-rank numbers) classified approximately 75% of the patients as being the same for their two mammograms.

### 5.4. Mammogram taxonomy

The best result was deemed to be the SONNET epoch with rank number 6 in the previous discussion. This result produced 39 stable classes, where the first was formed on the 2nd epoch of the run and the last was formed on the 84th epoch.

This result produced relatively narrow classes with an average within-class distance of 1.01 whilst retaining a typical average between-class distance of 2.00. Consequently, this result yielded a relatively high ratio in Figure 12 of 1.98. Approximately 37% of the patients had their two mammograms classified the same.

The chief features that discriminated class encodings were two textural features, namely angular second moment and contrast. Figures 14 to 19 are examples of the mammogram classes that were encoded by SONNET.

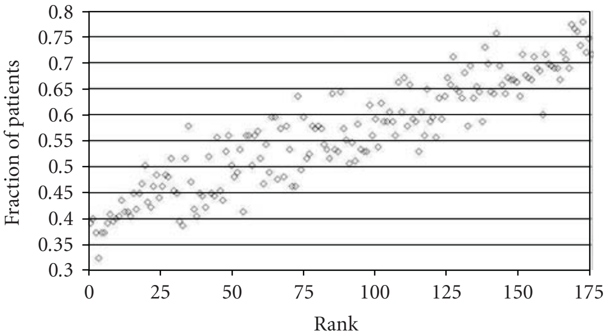


FIGURE 13: The fraction of patients whose two mammograms received the same classification for the best SONNET epochs. The epoch ranking corresponds to Figure 11.

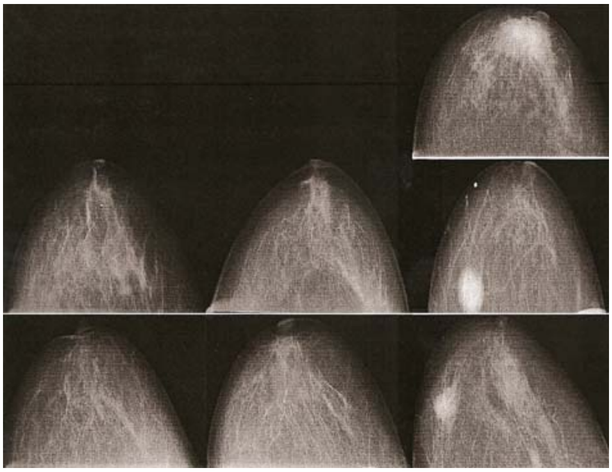


FIGURE 14: Example of a taxonomic class.

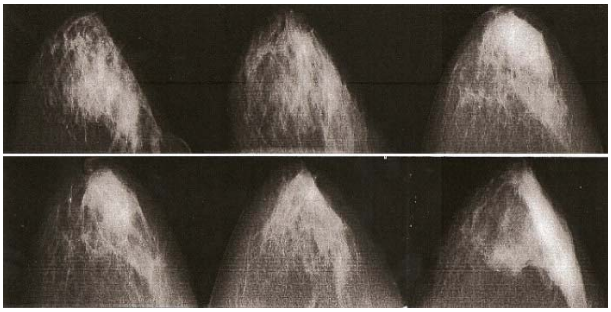


FIGURE 15: Example of a taxonomic class.

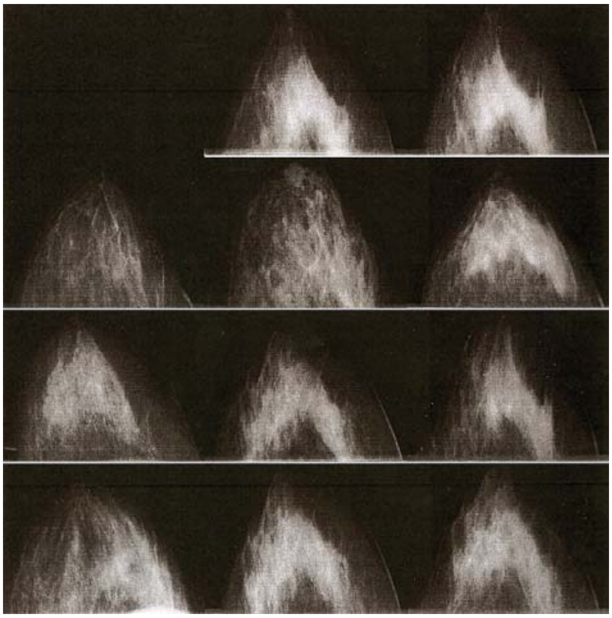


FIGURE 16: Example of a taxonomic class.

## 6. REFINING THE MAMMOGRAM TAXONOMY

### 6.1. Input feature selection

The main weakness with the current classification scheme is considered to be the manual specification of the image features which were extracted from the mammograms. The feature types and the regions from which they were extracted were designed to capture a priori knowledge about mammography, for instance, the importance of the retroareolar region. However, this manual specification necessarily requires arbitrary decisions, for instance, the quantitative position of the retroareolar region.

The image features and their corresponding region boundaries could be *automatically evolved* to produce an optimal input feature space. Two aspects of this are

(i) capturing *a priori* mammographic knowledge, for example, characteristic positions of lobular units, and

(ii) producing a high-quality feature space, for example, a minimal set of features with maximal orthogonality.

Other mammogram views could be used to extract input features in addition to the craniocaudal projection, for example, the mediolateral oblique view could be used.

### 6.2. Using control mammograms

Control mammograms could be exploited to refine the mammogram taxonomy. Control mammograms should be selected to represent clearly distinguishable mammogram classes. The classification scheme should initially be developed on these cases alone to shape classifier encodings. More ambiguous mammograms could then be introduced for subsequent classifier development. In order to achieve this, the classification scheme must be capable of *increhymental learning* and it must also address the so-called *stability-plasticity dilemma* [2]. SONNET satisfies these requirements.

### 6.3. Alternative classification schemes

A supervised classification scheme could be used to allow performance measures to actively drive classifier

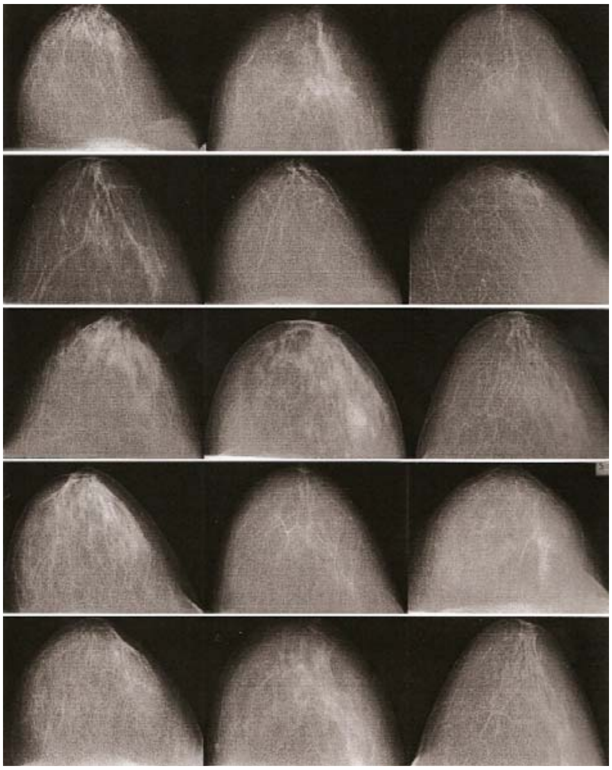


Figure 17: Example of a taxonomic class.

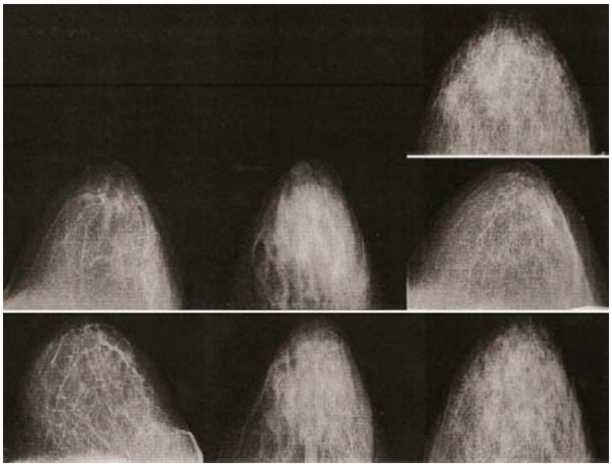


Figure 18: Example of a taxonomic class.

development in a manner consistent with the passive discovery of optimal SONNET classifications, as outlined in Section 4.5.

An *evolutionary computing* (EC) technique could form an alternative classification scheme. EC is a flexible and adaptable technique and consequently it could combine a number of the processing stages detailed in the above protocol for producing a mammogram taxonomy. For example, EC could optimize its own subset of input feature types by processing raw image features directly.

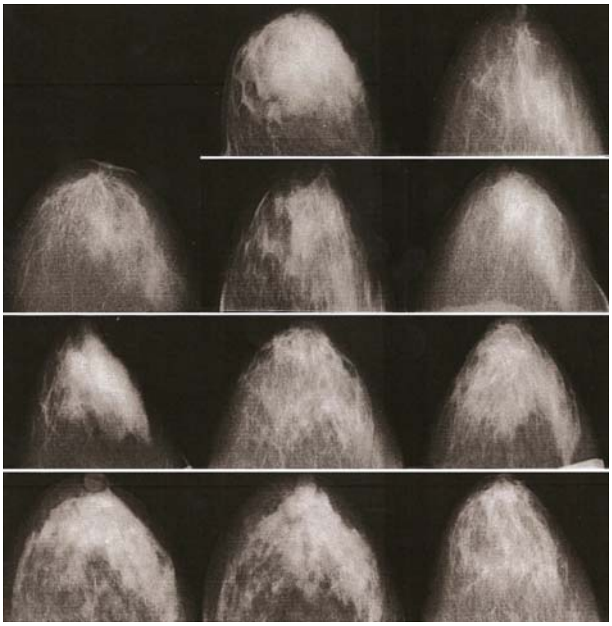


Figure 19: Example of a taxonomic class.

## 7. CONCLUSIONS

This study has developed a mammogram taxonomy by using an unsupervised classification scheme called SONNET. The encoded mammogram classes captured typical subtleties which discriminate mammograms. SONNET's controlling parameters were varied to govern the coarseness of the taxonomies. The developed classification scheme is considered to be a successful prototype but the scheme's efficacy is yet to be established.

The study shows promise for researching automated computational tools to assist with the detection of mammographic abnormalities. A mammogram taxonomy can be exploited to aid the detection of cancerous lesions via asymmetry identification [14], that is, by identifying anomalies between a patient's left and right mammograms. The evidence for cancerous lesions within the complex breast tissue can be very subtle, so mammogram features must capture localized information in a contextual manner, that is, *multiscale* features are required.

The authors have developed an evolutionary computation approach to discover multiscale features in imagery for a target detection application [15, 16]. This scheme used a *data crawler* which was evolved to gather evidence to discriminate target objects from nontarget objects. The crawler focused on low-level features in its immediate vicinity and processed these in the context of higher-level features collected over the crawler's trail.

As the data crawler has been developed for target detection in imagery, it is highly transferable to the problem of lesion detection in mammograms. The crawler could scrutinize mammogram areas which possess the greatest asymmetry and thus focus on candidate lesions. The evolutionary approach allows the crawler to discover its own multiscale features which best locate lesions.

The search for multiscale features over a diverse set of mammograms represents a very challenging problem, due to the high dimensionality of the potential search space. Hence, it is desirable to segregate the problem into multiple subproblems with less diversity. This can be achieved by exploiting the mammogram taxonomy as a preprocessing stage. This stage would classify a patient's mammograms, and thus would allow a data crawler to be evolved to specialize in only these taxonomic classes. Multiple crawlers could then be evolved, each of which specializes on its own subset of classes. Hence, the taxonomy would greatly constrain the search space in order to optimize asymmetry identification, and consequently, lesion detection.

## ACKNOWLEDGMENT

The views expressed are those of the authors and do not necessarily reflect those of QinetiQ. The authors are grateful to Professor Tabár for facilitating images in this study and for enthusing us to learn more about breast cancer screening.

## REFERENCES

[1] S. W. Duffy, R. A. Smith, R. Gabe, L. Tabár, A. M. Yen, and T. H. Chen, "Screening for breast cancer," *Surgical Oncology Clinics of North America*, vol. 14, no. 4, pp. 671–697, 2005.

[2] G. Carpenter and S. Grossberg, "ART2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, no. 23, pp. 4919–4930, 1987.

[3] L. Tabár, *Teaching Course of Mammography: Diagnosis and In-depth Differential Diagnosis of Breast Diseases*, Mammography Education, Cave Creek, Ariz, USA, 2006.

[4] D. Howard, S. C. Roberts, and L. Tabár, "Mammography taxonomy for improvement of lesion detection rates," in *Proceedings of the 6th International Workshop on Digital Mammography (IWDM '02)*, pp. 27–32, Bremen, Germany, June 2002.

[5] N. Jamal, K.-H. Ng, L.-M. Looi, et al., "Quantitative assessment of breast density from digitized mammograms into Tabar's patterns," *Physics in Medicine and Biology*, vol. 51, no. 22, pp. 5843–5857, 2006.

[6] S. Beer, *Brain of the Firm*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1981.

[7] N. F. Boyd, C. Wolfson, M. Moskowitz, et al., "Observer variation in the classification of mammographic parenchymal patterns," *Journal of Chronic Diseases*, vol. 39, no. 6, pp. 465–472, 1986.

[8] A. Nigrin, *Neural Networks for Pattern Recognition*, Bradford Books, MIT Press, Cambridge, Mass, USA, 1993.

[9] D. O. Hebb, *The Organization of Behavior*, John Wiley & Sons, New York, NY, USA, 1949.

[10] O. Paulsen and T. J. Sejnowski, "Natural patterns of activity and long-term synaptic plasticity," *Current Opinion in Neurobiology*, vol. 10, no. 2, pp. 172–179, 2000.

[11] S. Harford, "Automatic segmentation, learning and retrieval of melodies using a self-organizing neural network," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR '03)*, Baltimore, Md, USA, October 2003.

[12] J. A. Marshall and V. S. Gupta, "Generalization and exclusive allocation of creditin unsupervised category learning," *Network: Computation in Neural Systems*, vol. 9, no. 2, pp. 279–302, 1998.

[13] M. W. Spratling and M. H. Johnson, "Neural coding strategies and mechanisms of competition," *Cognitive Systems Research*, vol. 5, no. 2, pp. 93–117, 2004.

[14] D. Scutt, G. A. Lancaster, and J. T. Manning, "Breast asymmetry and predisposition to breast cancer," *Breast Cancer Research*, vol. 8, no. 2, article R14, 2006.

[15] D. Howard, S. C. Roberts, and C. Ryan, "Machine vision: exploring context with genetic programming," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '02)*, pp. 756–763, Morgan Kaufmann, New York, NY, USA, July 2002.

[16] D. Howard, S. C. Roberts, and C. Ryan, "Pragmatic genetic programming strategy for the problem of vehicle detection in airborne reconnaissance," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1275–1288, 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 346129, 8 pages
doi:10.1155/2008/346129

*Research Article*

# Validation of Alternating Kernel Mixture Method: Application to Tissue Segmentation of Cortical and Subcortical Structures

**Nayoung A. Lee, Carey E. Priebe, Michael I. Miller, and J. Tilak Ratnanather**

*Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218, USA*

Correspondence should be addressed to Nayoung A. Lee, nayoung@cis.jhu.edu

Received 4 September 2007; Revised 28 February 2008; Accepted 25 June 2008

Recommended by Daniel Howard

This paper describes the application of the alternating Kernel mixture (AKM) segmentation algorithm to high resolution MRI subvolumes acquired from a 1.5T scanner (hippocampus, $n = 10$ and prefrontal cortex, $n = 9$) and a 3T scanner (hippocampus, $n = 10$ and occipital lobe, $n = 10$). Segmentation of the subvolumes into cerebrospinal fluid, gray matter, and white matter tissue is validated by comparison with manual segmentation. When compared with other segmentation methods that use traditional Bayesian segmentation, AKM yields smaller errors ($P < .005$, exact Wilcoxon signed rank test) demonstrating the robustness and wide applicability of AKM across different structures. By generating multiple mixtures for each tissue compartment, AKM mimics the increased variation of manual segmentation in partial volumes due to the highly folded tissues. AKM's superior performance makes it useful for tissue segmentation of subcortical and cortical structures in large-scale neuroimaging studies.

## 1. INTRODUCTION

Current magnetic resonance image (MRI) studies investigate abnormalities of cortical and subcortical structures in neurodevelopmental and neurodegenerative disorders.These studies require a delineation of a region of interest (ROI) by manual segmentation by an expert rater. For example, studies on Alzheimer's disease and mild cognitive impairment examine the hippocampus [1] while those in schizophrenia have studied the occipital lobe and prefrontal cortex [2, 3]. Once the ROI is defined, segmentation into tissue types such as gray matter (GM), white matter (WM), or cerebrospinal fluid (CSF) can assess subtle volume changes caused by disease [4, 5]. While manual segmentation would provide gold standard, it is labor intensive limiting the number of subjects in any study [6]. Also, the rater needs to be trained to ensure small inter- or intrarater variation. Therefore, it is necessary to develop a method that allows for efficient processing of large number of subjects with high inter- or intrarater reliability, thereby increasing statistical power. Such a method will facilitate greater understanding of shape change in networks of cortical structures implicated in neuropsychiatric diseases [7, 8].

A variety of methods have been proposed for the segmentation of subcortical tissue such as the hippocampus [9–11] and cortical tissues such as prefrontal cortex, cingulate cortex, and planum temporale [12–16]. However, even though tissue classification methods have been improving in their performance, relatively low accuracy (comparing with expert-rater standards) has prevented accurate structural segmentation, for example, distinguishing the hippocampus from surrounding structures in the medial temporal lobe.

Partial volume voxels containing multiple tissue types present challenges to traditional Bayesian tissue classification methods [17–24] that model each tissue type as a fixed-bandwidth, single Gaussian in mixture-of-Gaussian models. Priebe et al. [25] proposed an alternating Kernel mixture (AKM) method which allowed for the flexibility of a Gaussian mixture model, with bandwidth, and the number of Gaussians selected adaptively from the data for each tissue type. The purpose of our study was to compare the performance of AKM and traditional Bayesian methods. The two methods were compared by determining which method was closer to the manual segmentation (ground truth) of cortical and subcortical structures in MRI subvolumes acquired from 1.5T and 3T scanners.

The manuscript is organized as follows. Section 2 describes the AKM mixture modeling methodology in detail and other Bayesian-based segmentation methods. Section 3 describes the dataset being investigated and image analysis employed. Section 4 reports the results.

## 2. METHOD

### 2.1. Alternating Kernel mixture method

Priebe and Marchette [26] and James et al. [27] introduced a semiparametric solution to the problem of estimating the common probability density function for multiple identically distributed random variables. Their solution is an iterative one that combines parametric and nonparametric estimates with a resulting model that incorporates both the complexity and the smoothness of the data.

We applied this method to the problem of MR segmentation. Gaussian mixture modeling is a popular segmentation technique. The marginal probability density function for the observations is

$$f = \sum_{c \in C} \pi_c f_c, \tag{1}$$

where $C := \{C, G, W\}$ is the set of tissue types (CSF, GM, and WM), $f_c$ are the class-conditional marginals, and $\pi_c$ are class-conditional mixing coefficients. These coefficients are nonnegative and sum to unity. Thus, the image is the sum of the three tissue types. Each class-conditional marginal is a mixture of normals given by

$$f_c = \sum_{t=1}^{k_c} \pi_{ct} \varphi_{ct}, \tag{2}$$

where $\pi_{ct}$ are the strictly positive, class-specific mixing coefficients, which sum to one, and $\varphi_{ct}$ are the Gaussian probabilities with a mean of $\mu_{ct}$ and a variance of $\sigma_{ct}^2$. Combining these equations we see that the marginals are given by

$$f = \sum_{c \in C} \pi_c \sum_{t=1}^{k_c} \pi_{ct} \varphi_{ct}. \tag{3}$$

The method estimates the class-conditional mixture complexities $k_c$, the mixing coefficients $\pi_c$, and the mixture components $\varphi_{ct}$. The Expectation-Maximization (EM) algorithm is used to estimate the means and variances of the components [18, 19]. The mixture complexities are estimated from the data.

The method alternates between parametric finite mixture estimates and nonparametric Kernel estimates. Each estimate is based on the previous one of the opposite type. The first step of the algorithm is to find a parametric estimate and a nonparametric estimate of the data. Then, at each iteration, a parametric estimate that minimizes the distance between the two previous estimates is computed. Using the parameter estimates thus derived, a nonparametric estimate is found. This continues until the distance between two consecutive parametric estimates is smaller than a desired constant.

Table 1: Voxel classification based on likelihood ratio test.

| Case | Classification |
|---|---|
| $r_1(x) > 1$ | Voxel labeled $C$ |
| $r_2(x) < 1$ | Voxel labeled $W$ |
| $r_1(x) < 1$ and $r_2(x) > 1$ | Voxel labeled $G$ |
| $r_1(x) > 1$ and $r_2(x) < 1$ | Should not occur |
| Tie | Determined arbitrarily |

The filtered Kernel estimate (i.e., the nonparametric estimate), with bandwidth $b$, is

$$\tilde{f}(x; X) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{k} \frac{\pi_t \varphi_t(X_i)}{f(X_i) b \sigma_t} \varphi_0\left(\frac{x - X_i}{b \sigma_t}\right), \tag{4}$$

where $X = \{X_1, \ldots, X_n\}$ is the subject's MR voxel observation, $\sigma_t^2$ is the variance of the $t$th component of the mixture, and $\varphi_0$ is the standard normal with zero mean and unit variance. The nonparametric estimates are each based on the parametric estimate from the previous iteration and are given by

$$\hat{f}^k = \arg\min_{f \in F^k} \| f - \tilde{f}^{k-1} \|_2^2, \tag{5}$$

where $F^k$ is the class of $k$-component Gaussian mixtures, and

$$\| f - g \|_2^2 := \int_{-\infty}^{\infty} (f(x) - g(x))^2 dx \tag{6}$$

is the integrated squared error.

To actually classify voxels, the Bayes plug-in classifier is used:

$$g(x) = \arg\max_{c \in C} \pi_c f_c(x), \tag{7}$$

where $x$ is the voxel to be labeled. The label is assigned to a class based on which one maximizes posterior probability of class membership. This can also be seen as a likelihood ratio test procedure given by

$$\begin{aligned} \mathrm{LRT}_{C/G(x)} &= \frac{\pi_C f_C(x)}{\pi_G f_G(x)} =: r_1(x), \\ \mathrm{LRT}_{G/W(x)} &= \frac{\pi_G f_G(x)}{\pi_W f_W(x)} =: r_2(x). \end{aligned} \tag{8}$$

Tissues are then classified according to Table 1.

This method results in the voxels being classified into three categories. Priebe et al. [25] showed how a training set could be used to determine the number of components for each tissue. However, the focus of this paper is on how this could be done on a case-by-case basis using visual inspection. It was found that two or three components of CSF, GM, and WM produced the best result; in a couple of cases the complexity was better modeled with four components.

### 2.2. Bayesian segmentation

For comparison, voxels are classified into three tissue types by Bayesian segmentation:

$$p(I_n \mid \mu_n(h_n), \sigma_n^2(h_n)) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2(h_n)}} e^{(-(I_n-\mu_n(h_n))^2/2\sigma_n^2(h_n))}, \tag{9}$$

where $I_n$ is the image intensity, $h_n$ is the anatomical label, $\mu_n$ is the mean, and $\sigma_n^2$ is the variance of the Gaussian density. The algorithm is

$$h_n = \arg\max_{h_n \in H} \sum_{n=1}^{N} \left( -\frac{1}{2}\log 2\pi\sigma_n^2(h_n) - \frac{1}{2}\frac{(I_n - \mu_n(h_n))^2}{\sigma_n^2(h_n)} + \log \pi(h_n) \right), \tag{10}$$

where $\pi(h_n)$ is the prior distribution that represents the relative amount of each of the tissue types and $H := \{C, G, W\}$. As with AKM, the EM algorithm is used to estimate the means and variances of the three tissues [18, 19].

### 2.3. Neyman-Pearson recalibration

Bayesian segmentation can be extended to two additional classes for $C/G$ and $G/W$ partial volumes which are optimally determined by [28]

$$\frac{p(I_n \mid h_n = G)}{p(I_n \mid h_n = \mathrm{CSF})} \underset{C}{\overset{G}{\gtrless}} \theta_{C/G}, \qquad \frac{p(I_n \mid h_n = W)}{p(I_n \mid h_n = G)} \underset{G}{\overset{W}{\gtrless}} \theta_{G/W} \tag{11}$$

at each voxel. Here, the four thresholds $(\theta_1, \ldots, \theta_4)$ are determined by the five Gaussians. Thresholds are selected to minimize the misclassification error (Section 3.5) such that $\theta_{C/G} = \theta_1 + t_{C/G}(\theta_2 - \theta_1)$ and $\theta_{G/W} = \theta_3 + t_{G/W}(\theta_4 - \theta_3)$, where $t_{C/G} \in [0, 1]$ and $t_{G/W} \in [0, 1]$. The means then are used to recalibrate the segmentations yielding new thresholds. This is referred to as Neyman-Pearson recalibration.

## 3. VALIDATION

### 3.1. Data acquisition

Four different sets of ROIs were extracted from subjects scanned via the magnetization prepared rapid gradient echo sequence on different scanners. Two came from a 3T scanner (10 hippocampi [29] and 5 pairs of left and right occipital lobes [30]); two came from a 1.5T scanner (10 hippocampi [31] and 9 prefrontal cortices [31]). Processed datasets were reformatted to 8 bit and interpolated to 1 × 1 × 1 $\mathrm{mm}^3$ isotropic voxels except for the prefrontal set with a resolution of 0.5 × 0.5 × 0.5 $\mathrm{mm}^3$, and are available at http://www.cis.jhu.edu/data.sets/index.html.

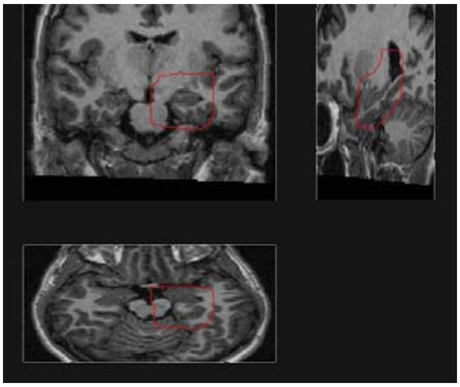


Figure 1: Hippocampus ROI mask delineated in red.

### 3.2. MRI subvolumes

To obtain a smaller ROI around a hippocampus, we manually outlined hippocampus and dilated it by 3 × 3 × 3 $\mathrm{mm}^3$ cubes with three iterations to generate a mask via BLOX (http://sourceforge.net/projects/blox/). Figure 1 shows an example of the mask generated for a left hippocampus. The prefrontal cortex [31] and occipital lobe [30] subvolumes were defined by an expert neuroanatomist.

### 3.3. Manual tissue segmentation

The 39 subvolumes were hand segmented into CSF, GM, and WM tissue compartments by three different raters in independent studies (e.g., [30, 31]) and blind to the autosegmentation. Segmentation was done by visual inspection on contiguous sagittal slices on Analyze software [32] and saved as Analyze image data with labels for CSF, GM, and WM.

### 3.4. Automated tissue segmentation

AKM and Bayesian segmentation were applied to the 39 subvolumes. For comparison, FreeSurfer [33] and BrainVoyager [34] were used to segment the hippocampi and occipital lobes, respectively. Neyman-Pearson segmentation was applied to prefrontal cortex. The EM algorithm [18] ensured that computations were done in real time.

The 10 hippocampus subvolumes from the 1.5T scanner were processed by FreeSurfer [33] to segment and label the volume by its anatomical structure. Each voxel was classified by a given anatomical label (i.e., hippocampus, ventricles). Then we group the structures into WM, GM, and CSF to create WM, GM, and CSF masks. Lateral ventricle and left inferior lateral ventricle were categorized as CSF. Cerebellum-exterior, hippocampus and amygdala were grouped as GM and cerebral white matter, thalamus proper, putamen, ventral diencephalon, and WM hypointensities were grouped as WM.

Table 2: Classification error for Bayesian and AKM for hippocampi (3T).

| | Bayesian | AKM |
|---|---|---|
| 1 | 0.106 | 0.099 |
| 2 | 0.124 | 0.107 |
| 3 | 0.234 | 0.183 |
| 4 | 0.133 | 0.120 |
| 5 | 0.283 | 0.114 |
| 6 | 0.186 | 0.103 |
| 7 | 0.153 | 0.131 |
| 8 | 0.133 | 0.126 |
| 9 | 0.273 | 0.172 |
| 10 | 0.163 | 0.153 |

Table 3: Classification error for Bayesian, AKM, and Neyman-Pearson for prefrontal cortices (1.5T).

| | Bayesian | Neyman-Pearson | AKM |
|---|---|---|---|
| 1 | 0.138 | 0.144 | 0.093 |
| 2 | 0.103 | 0.103 | 0.101 |
| 3 | 0.087 | 0.088 | 0.081 |
| 4 | 0.091 | 0.097 | 0.081 |
| 5 | 0.135 | 0.135 | 0.097 |
| 6 | 0.093 | 0.098 | 0.088 |
| 7 | 0.095 | 0.103 | 0.093 |
| 8 | 0.127 | 0.127 | 0.106 |
| 9 | 0.091 | 0.091 | 0.085 |

### 3.5. Quantification of segmentation accuracy

Segmentations were compared via the $L_1$ distance between two distributions as a measure of misclassification error. A cost is assigned to each labeled voxel. If it was labeled correctly, that cost is 0, and if labeled incorrectly, that cost is generally 1. This cost, called the $L_1$ distance, is

$$L_1 = \frac{1}{2N}\sum_{n=1}^{N}\sum_{i=1}^{m} \left| p^A(h_n = H_i \mid I_n) - p^M(h_n = H_i \mid I_n) \right|, \tag{12}$$

where $p^A(h_n \mid I_n)$ is the posteriori probability of hypothesis $h_n$ at voxel $n$ for the automated, $p^M(h_n \mid I_n)$ is the same for the manual segmentation, and $m$ is the number of tissue types [18, 19, 28, 35]. The distance measures agreement between segmentations based on distance between probability distributions [36]. The standard overlap measures penalize small objects assuming that most of the error occurs at the boundary of objects thus $L_1$ distance is more appropriate for assessing 3D segmentation [37]. Another standard measure, the Dice measure, was also used [38].

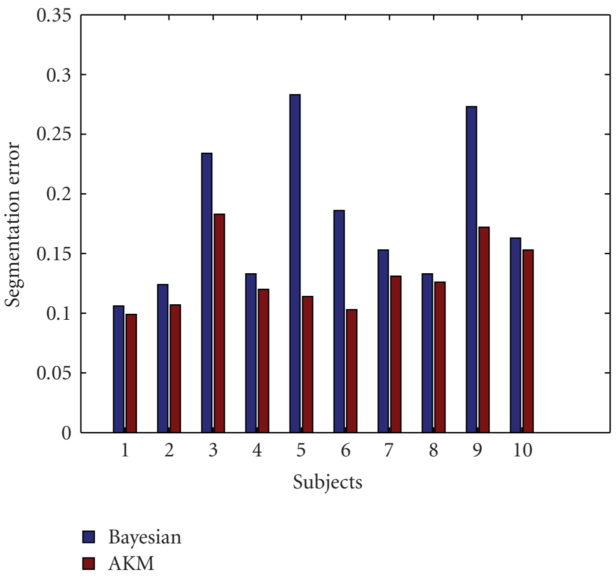


Figure 2: Classification error for Bayesian (blue) and AKM (red) for hippocampi (3T).

Table 4: Classification error for Bayesian, BrainVoyager, and AKM for occipital lobes (3T).

| | Bayesian | BrainVoyager | AKM |
|---|---|---|---|
| 1 | 0.170 | 0.199 | 0.119 |
| 2 | 0.149 | 0.202 | 0.093 |
| 3 | 0.243 | 0.221 | 0.099 |
| 4 | 0.210 | 0.202 | 0.096 |
| 5 | 0.236 | 0.244 | 0.112 |
| 6 | 0.224 | 0.237 | 0.128 |
| 7 | 0.165 | 0.111 | 0.099 |
| 8 | 0.224 | 0.117 | 0.104 |
| 9 | 0.157 | 0.248 | 0.119 |
| 10 | 0.146 | 0.237 | 0.121 |

Table 5: Classification error for Bayesian, FreeSurfer and AKM for ten hippocampi (1.5T).

| | Bayesian | FreeSurfer | AKM |
|---|---|---|---|
| 1 | 0.121 | 0.145 | 0.113 |
| 2 | 0.162 | 0.225 | 0.161 |
| 3 | 0.110 | 0.144 | 0.096 |
| 4 | 0.131 | 0.178 | 0.109 |
| 5 | 0.175 | 0.191 | 0.146 |
| 6 | 0.129 | 0.169 | 0.119 |
| 7 | 0.121 | 0.190 | 0.121 |
| 8 | 0.121 | 0.165 | 0.121 |
| 9 | 0.155 | 0.196 | 0.139 |
| 10 | 0.128 | 0.143 | 0.120 |

## 4. RESULTS

Tables 2, 3, 4, and 5 and Figures 2, 3, 4, and 5 show that $L_1$ distances for AKM method are lower than those

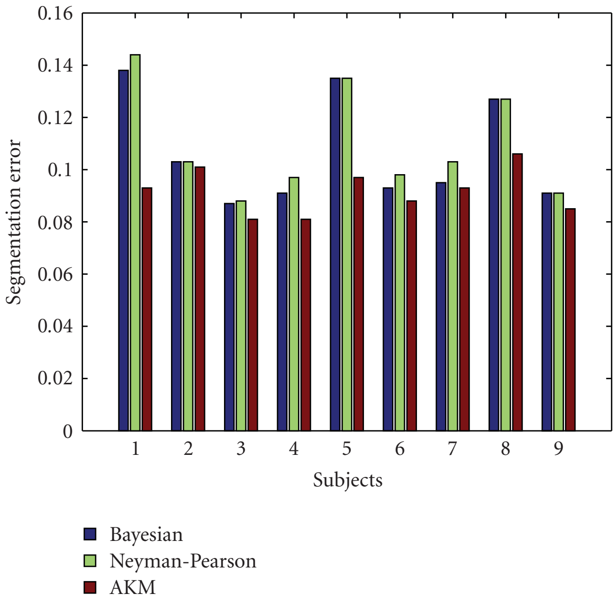


FIGURE 3: Classification error for Bayesian (blue), Neyman-Pearson (green), and AKM (red) for prefrontal cortices (1.5T).

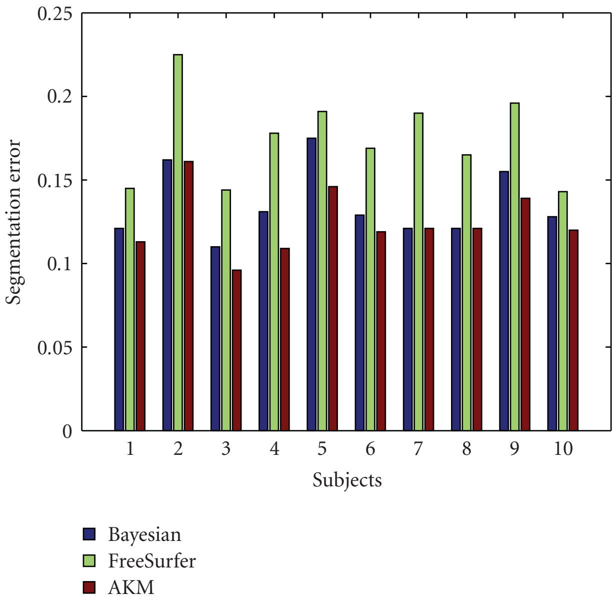


FIGURE 5: Classification error for Bayesian (blue), FreeSurfer (green), and AKM (red) for ten hippocampi (1.5T).

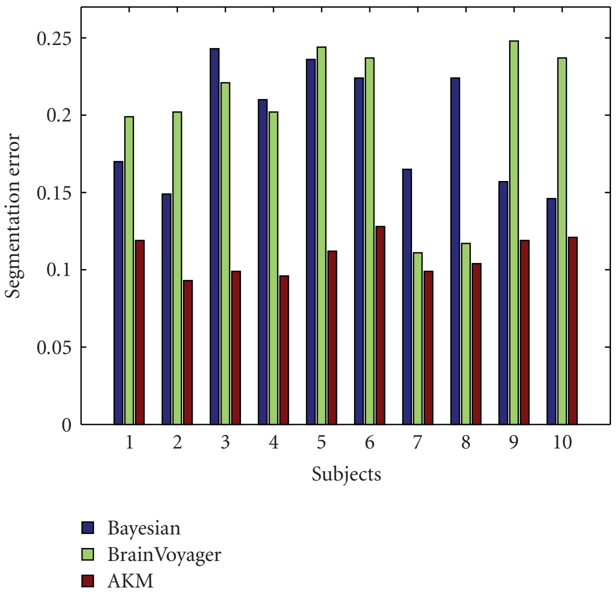


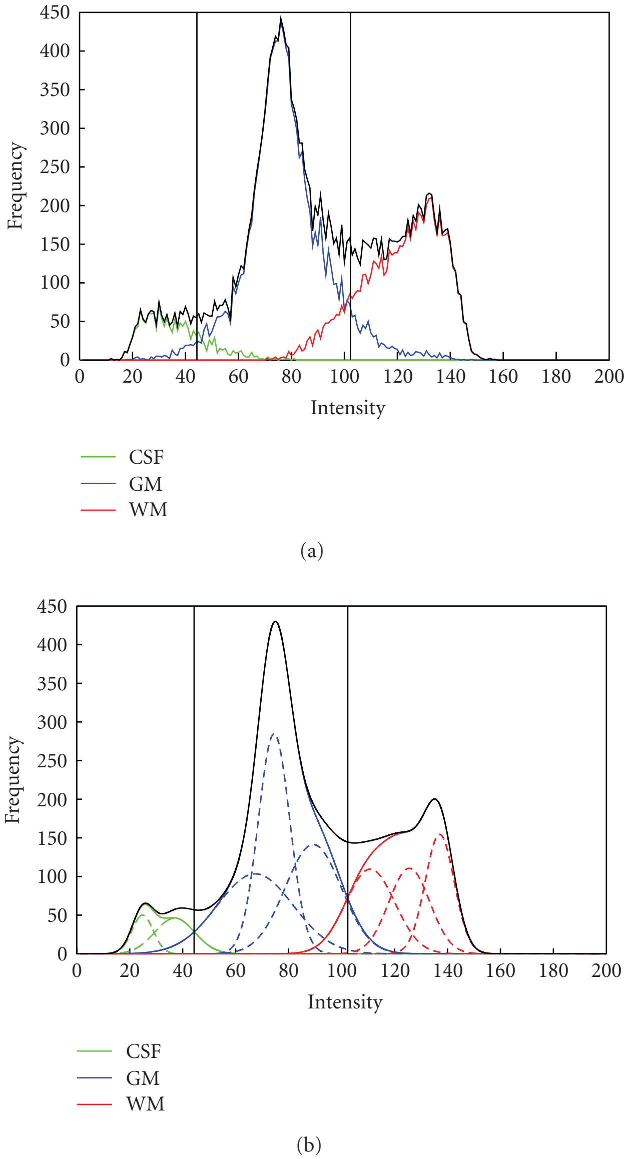
FIGURE 4: Classification error for Bayesian (blue), BrainVoyager (green), and AKM (red) for occipital lobes (3T).

for Bayesian and other segmentation methods ($P < .005$, Exact Wilcoxon signed rank test). Lower $L_1$ distances mean that AKM segmentation have more overlap with manual segmentation than other methods. Dice measures for AKM were consistently smaller than other methods.

Figures 6 and 7 explain the reason for low classification errors of AKM. Green, red, and blue curves show the intensity profile of voxels labeled as CSF, GM, and WM, respectively. The figures show how intensity histograms for manual segmentation voxels are similar to those for AKM segmentation. Vertical lines are threshold intensity values calculated from AKM method. Manual segmentation histograms show that each tissue type has wide range of intensities thus resulting in large overlaps between tissue types due to partial volume problems where the boundaries between tissue types are not obvious. Figure 7 shows how the large tails for each tissue types is captured by AKM yielding more accurate threshold values compared with the single Gaussian approach. Further, Table 3 and Figure 3 shows that AKM models the partial volume better than Neyman-Pearson; note that Neyman-Pearson yielded larger errors than Bayesian since the recalibration was based on the averaged thresholds. Finally, Figures 8, 9, and 10 show views of the AKM segmentation of hippocampus, prefrontal cortex, and occipital lobe subvolumes.

## 5. CONCLUSION

This paper describes an algorithm that models each tissue type in brain MRI subvolumes as a semiparametric mixture of Gaussians. The classification method which uses this algorithm results in better segmentation than a traditional, single-component Bayesian method especially when there is not enough CSF or WM in the subvolume. Human raters are good at segmenting partial volume voxels by adapting to the high variance of intensities in these regions. AKM is also able to adaptively select the bandwidth and the number of Gaussians for each tissue type. Thus, AKM approximated the manual segmentation more closely compared to Bayesian methods. AKM can automatically delineate cortical and subcortical structures which can be distinguished by intensity information. However, there are structures that cannot be segmented by intensity alone. For example,



(a)

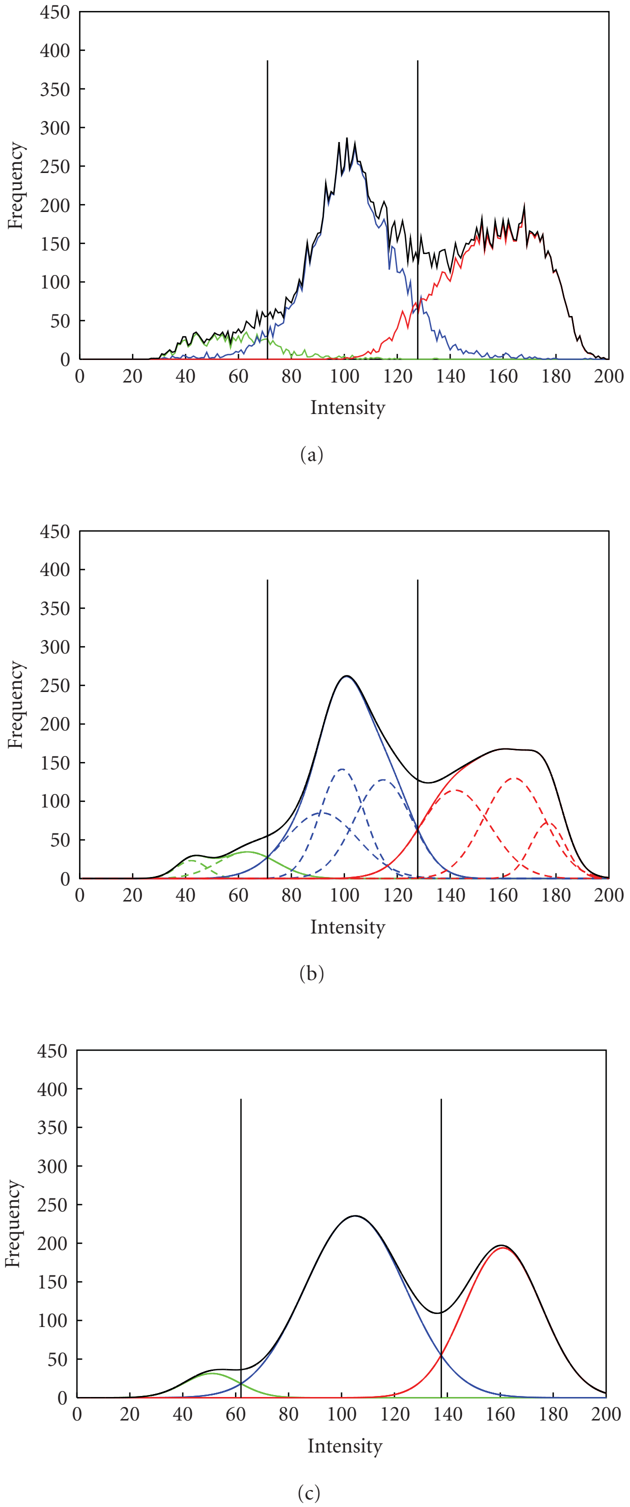
(b)

Figure 6: Intensity histogram for hippocampus: hand (a) and AKM (b) segmentation with vertical lines from AKM threshold. Green, blue, and red correspond to CSF, GM, and WM segmented voxels, respectively.

anterior boundary of the hippocampus merges with the amygdala which has similar intensity [39] or the anatomical boundary of prefrontal cortex and occipital lobe has to be defined with spatial information. For these structures, AKM may be useful when combined with mapping and image registration approach. Also, AKM can be applied to other imaging modalities of other anatomical structures, such as segmenting myocardium, blood, and bone in a cardiac CT scan.

## ACKNOWLEDGMENTS

This work was supported by National Institute of Health Grant nos. (R01 EB00975-01, P41 RR15241-01A1, P41 RR15241-02S1, P50 MH071616, and P01 AG026276). Assistance from Dr. L. Wang, Dr. Y. Park, Dr. A. Qiu, Dr. S. Yantis, N. Penumetcha, and M. Hosakere is appreciated.



(a)

(b)

(c)

Figure 7: Intensity histogram for occipital lobe: hand- (a), AKM (b), and Bayesian segmentation (c) with vertical lines from AKM ((a) and (b)) and Bayesian (c).

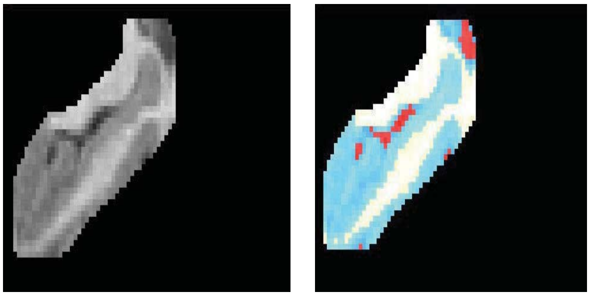


Figure 8: Sagittal view of left hippocampus. (a) MRI. (b) AKM segmentation (blue-GM, white-WM, red-CSF).

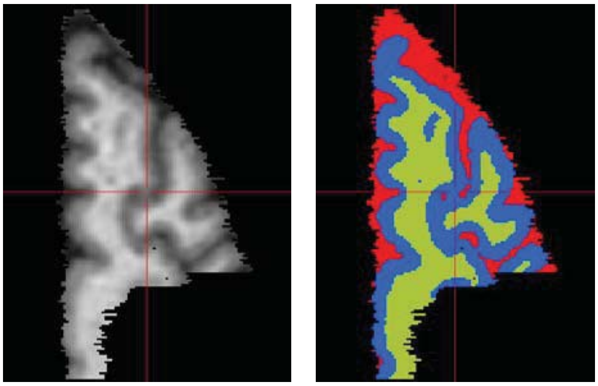


Figure 9: Axial view of prefrontal cortex. (a) MRI. (b) AKM segmentation (blue-GM, green-WM, red-CSF).

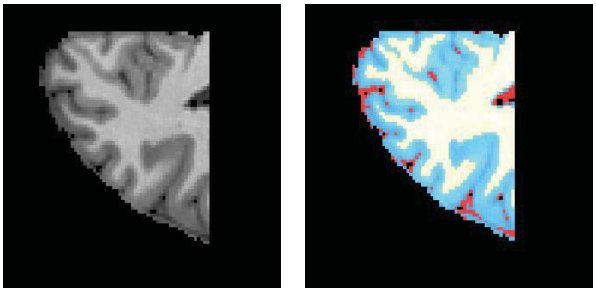


Figure 10: Axial view of left occipital lobe. (a) MRI. (b) AKM segmentation (blue-GM, white-WM, red-CSF).

## REFERENCES

[1] E. Geuze, E. Vermetten, and J. D. Bremner, "MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed," *Molecular Psychiatry*, vol. 10, no. 2, pp. 147–159, 2005.

[2] N. C. Andreasen, "Linking mind and brain in the study of mental illnesses: a project for a scientific psychopathology," *Science*, vol. 275, no. 5306, pp. 1586–1593, 1997.

[3] T. Onitsuka, R. W. McCarley, N. Kuroki, et al., "Occipital lobe gray matter volume in male patients with chronic schizophrenia: a quantitative MRI study," *Schizophrenia Research*, vol. 92, no. 1–3, pp. 197–206, 2007.

[4] R. E. Gur, B. I. Turetsky, P. E. Cowell, et al., "Temporolimbic volume reductions in schizophrenia," *Archives of General Psychiatry*, vol. 57, no. 8, pp. 769–775, 2000.

[5] Y. Hirayasu, S. Tanaka, M. E. Shenton, et al., "Prefrontal gray matter volume reduction in first episode schizophrenia," *Cerebral Cortex*, vol. 11, no. 4, pp. 374–381, 2001.

[6] R. Kikinis, M. E. Shenton, G. Gerig, et al., "Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging," *Journal of Magnetic Resonance Imaging*, vol. 2, no. 6, pp. 619–629, 1992.

[7] J. G. Csernansky, L. Wang, D. Jones, et al., "Hippocampal deformities in schizophrenia characterized by high dimensional brain mapping," *American Journal of Psychiatry*, vol. 159, no. 12, pp. 2000–2006, 2002.

[8] L. Wang, D. Y. Lee, E. Bailey, et al., "Validity of large-deformation high dimensional brain mapping of the basal ganglia in adults with Tourette syndrome," *Psychiatry Research*, vol. 154, no. 2, pp. 181–190, 2007.

[9] S. Bouix, J. C. Pruessner, D. Louis Collins, and K. Siddiqi, "Hippocampal shape analysis using medial surfaces," *NeuroImage*, vol. 25, no. 4, pp. 1077–1089, 2005.

[10] J. W. Haller, A. Banerjee, G. E. Christensen, et al., "Three-dimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas," *Radiology*, vol. 202, no. 2, pp. 504–510, 1997.

[11] O. T. Carmichael, H. A. Aizenstein, S. W. Davis, et al., "Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 27, no. 4, pp. 979–990, 2005.

[12] J. T. Ratnanather, K. N. Botteron, T. Nishino, et al., "Validating cortical surface analysis of medial prefrontal cortex," *NeuroImage*, vol. 14, no. 5, pp. 1058–1069, 2001.

[13] A. Qiu, M. Vaillant, P. Barta, J. T. Ratnanather, and M. I. Miller, "Surface-based Gaussian random field model with application to cortical thickness variation of left planum temporale in schizophrenia and bipolar disorder," *Human Brain Mapping*, vol. 29, no. 8, pp. 923–925, 2008.

[14] A. Qiu, L. Younes, L. Wang, et al., "Combining anatomical manifold information via diffeomorphic metric mappings for studying cortical thinning of the cingulate gyrus in schizophrenia," *NeuroImage*, vol. 37, no. 3, pp. 821–833, 2007.

[15] L. Wang, M. Hosakere, J. C. L. Trein, et al., "Abnormalities of cingulate gyrus neuroantomy in schizophrenia," *Schizophrenia Research*, vol. 93, no. 1–3, pp. 66–78, 2007.

[16] M. I. Miller, M. Hosakere, A. R. Barker, et al., "Labeled cortical mantle distance maps of the cingulate quantify differences between dementia of the Alzheimer type and healthy aging," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, pp. 15172–15177, 2003.

[17] B. Fischl, A. Liu, and A. M. Dale, "Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 70–80, 2001.

[18] M. Joshi, J. Cui, K. Doolittle, et al., "Brain segmentation and the generation of cortical surfaces," *NeuroImage*, vol. 9, no. 5, pp. 461–476, 1999.

[19] M. I. Miller, A. B. Massie, J. T. Ratnanather, K. N. Botteron, and J. G. Csernansky, "Bayesian construction of geometrically based cortical thickness metrics," *NeuroImage*, vol. 12, no. 6, pp. 676–687, 2000.

[20] W. M. Wells III, W. E. L. Crimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Transactions on Medical Imaging*, vol. 15, no. 4, pp. 429–442, 1996.

[21] Z. Y. Shan, G. H. Yue, and J. Z. Liu, "Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images," *NeuroImage*, vol. 17, no. 3, pp. 1587–1598, 2002.

[22] C. A. Cocosco, A. P. Zijdenbos, and A. C. Evans, "A fully automatic and robust brain MRI tissue classification method," *Medical Image Analysis*, vol. 7, no. 4, pp. 513–527, 2003.

[23] T. J. Grabowski, R. J. Frank, N. R. Szumski, C. K. Brown, and H. Damasio, "Validation of partial tissue segmentation of single-channel magnetic resonance images of the brain," *NeuroImage*, vol. 12, no. 6, pp. 640–656, 2000.

[24] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, 2001.

[25] C. E. Priebe, M. I. Miller, and J. T. Ratnanather, "Segmenting magnetic resonance images via hierarchical mixture modelling," *Computational Statistics & Data Analysis*, vol. 50, no. 2, pp. 551–567, 2006.

[26] C. E. Priebe and D. J. Marchette, "Alternating kernel and mixture density estimates," *Computational Statistics & Data Analysis*, vol. 35, no. 1, pp. 43–65, 2000.

[27] L. F. James, C. E. Priebe, and D. J. Marchette, "Consistent estimation of mixture complexity," *The Annals of Statistics*, vol. 29, no. 5, pp. 1281–1296, 2001.

[28] J. T. Ratnanather, L. Wang, M. B. Nebel, et al., "Validation of semiautomated methods for quantifying cingulate cortical metrics in schizophrenia," *Psychiatry Research: Neuroimaging*, vol. 132, no. 1, pp. 53–68, 2004.

[29] C. B. Kirwan, C. K. Jones, M. I. Miller, and C. E. L. Stark, "High-resolution fMRI investigation of the medial temporal lobe," *Human Brain Mapping*, vol. 28, no. 10, pp. 959–966, 2006.

[30] A. Qiu, B. J. Rosenau, A. S. Greenberg, et al., "Estimating linear cortical magnification in human primary visual cortex via dynamic programming," *NeuroImage*, vol. 31, no. 1, pp. 125–138, 2006.

[31] J. P. John, L. Wang, A. J. Moffitt, H. K. Singh, M. H. Gado, and J. G. Csernansky, "Inter-rater reliability of manual segmentation of the superior, inferior and middle frontal gyri," *Psychiatry Research*, vol. 148, no. 2-3, pp. 151–163, 2006.

[32] R. A. Robb, D. P. Hanson, R. A. Karwoski, A. G. Larson, E. L. Workman, and M. C. Stacy, "Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis," *Computerized Medical Imaging and Graphics*, vol. 13, no. 6, pp. 433–454, 1989.

[33] B. Fischl, D. H. Salat, E. Busa, et al., "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[34] N. Kriegeskorte and R. Goebel, "An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes," *NeuroImage*, vol. 14, no. 2, pp. 329–346, 2001.

[35] J. T. Ratnanather, P. E. Barta, N. A. Honeycutt, et al., "Dynamic programming generation of boundaries of local coordinatized submanifolds in the neocortex: application to the planum temporale," *NeuroImage*, vol. 20, no. 1, pp. 359–377, 2003.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[37] G. Gerig, M. Jomier, and M. Chakos, "Valmet: a new validation tool for assessing and improving 3D object segmentation," in *Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '01)*, vol. 2208 of *Lecture Notes in Computer Science*, pp. 516–523, Utrecht, The Netherlands, October 2001.

[38] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: method and validation," *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

[39] M. A. Munn, J. Alexopoulos, T. Nishino, et al., "Amygdala volume analysis in female twins with major depression," *Biological Psychiatry*, vol. 62, no. 5, pp. 415–422, 2007.