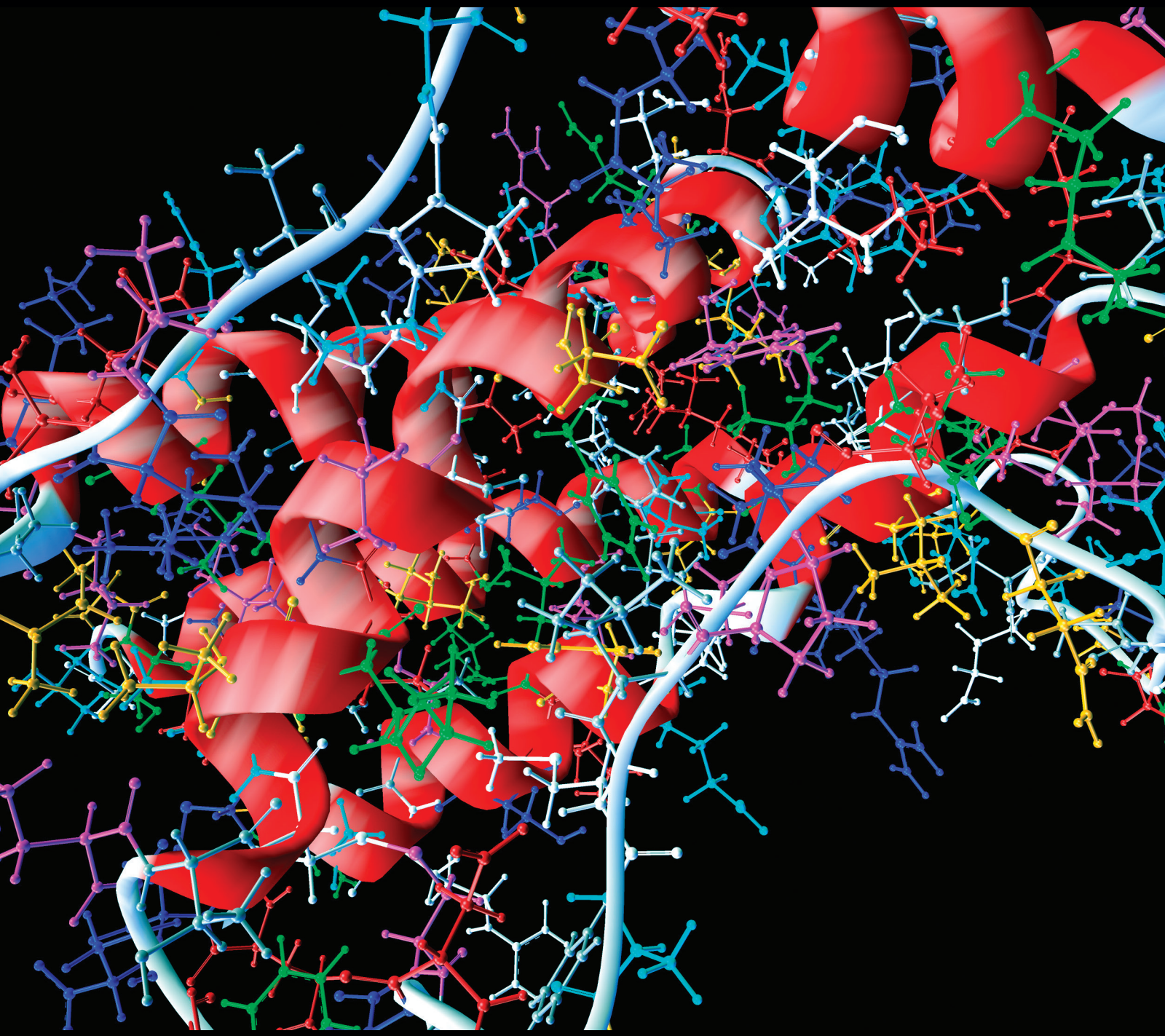
Computational and Mathematical Methods in Medicine

Integrated Approach in Systems Biology

Guest Editors: Huiru Zheng, Rui Jiang, and Zhongming Zhao



Integrated Approach in Systems Biology

Computational and Mathematical Methods in Medicine

Integrated Approach in Systems Biology

Guest Editors: Huiru Zheng, Rui Jiang, and Zhongming Zhao

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 656473, 2 pages
http://dx.doi.org/10.1155/2014/656473

Editorial

Integrated Approach in Systems Biology

Huiru Zheng,[1] Rui Jiang,[2] and Zhongming Zhao[3]

[1]*Computer Science Research Institute, School of Computing and Mathematics, University of Ulster, Shore Road, Newtownabbey, County Antrim BT37 0QB, UK*
[2]*Bioinformatics Division, Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China*
[3]*Departments of Biomedical Informatics, Psychiatry, and Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*

Correspondence should be addressed to Huiru Zheng; h.zheng@ulster.ac.uk

Received 13 October 2014; Accepted 13 October 2014; Published 31 December 2014

Systems biology is a field within biology, aiming at understanding biological processes at the systems level and emerging from dynamic interactions of individual components that operate at multiple spatiotemporal scales. It is now an established and fundamental interdisciplinary research field. Systems biology studies biological systems by systematically perturbing them (biologically, genetically, chemically, or other); monitoring the gene, protein, metabolite, and informational pathway responses; integrating these data; ultimately formulating mathematical models that describe the structure of the system and predict its response to individual perturbations. Integrated "omics" (such as genome-wide measurements of transcripts, protein levels, or metabolite level) approaches have created exciting opportunities for systems biology and other biological researches thanks to the rapid advancement of high throughput biotechnologies. Computational methods such as data preprocessing, representation, modeling, measurement, interpretation, prediction, visualisation, and simulation have been well applied to understand biological processes and biological systems.

We organised this special issue to provide an international forum to discuss the most recent developments in the field regarding integrated data analysis approaches in systems biology research such as pattern recognition and prediction, modeling and simulation, and data representation and visualisation. This special issue featured "integrated approach" and "complex biological system" themes. We are interested in both new theories and tools in this area as well as their applications in systems biology. The potential topics include (i) large-scale or cross-species data integration for the reconstruction of networks and pathways; (ii) genomic data analysis using systems biology approaches; (iii) quantitative understanding of the dynamics of regulatory, signaling, interaction, and metabolic networks through modeling and simulation techniques; (iv) prediction of protein/RNA structure and biological network-based interactions; (v) data integration and knowledge-driven approach in biomarker identification and drug discovery; (vi) enhancement and enablement of knowledge discovery in functional genomics of disease and other phenotypes through integrated omics approach; (vii) semantic webs and ontology-driven biological data integration methods; (viii) development of integrated systems biology visualisation and analysis tools; (ix) development of integrated systems biology visualisation and analysis tools; (x) integrating approaches in transitional bioinformatics and personalized medicine.

Eight manuscripts were submitted in response to this special issue and six were finally accepted for publication, ranging from mathematical model, computational pipeline, and engineering design to computational models, with the applications at molecular, neuronal, protein, gene regulatory network, and gene ontology levels.

The comparison on biology sequences is one of the most important tasks in analyzing similarities of function and properties. W. Deng and Y. Luan integrated the dual-vector curve (DV-curve) and the detailed hydrophobic-hydrophilic (HP) model of amino acids, in the representation and visualization of protein sequences. Although the information

might be lost in the representation, their results showed that the proposed method is efficient and feasible when focusing on the important part of the sequences.

L. Pasotti and S. Zucca reviewed the recent advances and computational tools in biological engineering design on which predictability issues in promoters, ribosome binding sites, coding sequences, transcriptional terminators, and plasmids were specifically discussed. The authors suggested that bottom-up approaches are urgently needed in order to refine and exploit the full potential of synthetic biology and a mixture of prediction tools could rapidly boost the efficiency of biological engineering by providing a smaller search space than fully random-based approaches.

In "*A pipeline for neuron reconstruction based on spatial sliding volume filter seeding*," D. Sui et al. proposed a pipeline with a new seeding method for the construction of neuron structures from three-dimensional microscopy images stacks, which will be beneficial to three- dimensional neuron reconstruction and detection.

Gene regulatory networks consist of interactions between large number of genes and their regulators and are involved in every biological process. L. P. Tian et al. designed a state observer to estimate the states of genetic regulatory networks with time delays from available measurements. Furthermore, based on linear matrix inequality approach, a gene repressillatory network was employed to illustrate the effectiveness of the proposed design approach.

In "*Effects of maximal sodium and potassium conductance on the stability of Hodgkin-Huxley model*," Y. Zhang et al. applied stability theory in the model design to investigate the importance of maximal sodium conductance and maximal potassium conductance. The study could help in researches relevant to diseases caused by maximal conductance anomaly.

In "*Correlating information contents of gene ontology terms to infer semantic similarity of gene products*," the author proposed a new semantic gene ontology similarity measurement. A gene product was represented as a vector that is composed of information contents of gene ontology terms annotated for the gene product, and the pairwise similarity between two gene products was viewed as the relatedness of their corresponding vectors using three measures: Pearson's correlation coefficient, cosine similarity, and Jaccard index.

Acknowledgments

We are grateful to the anonymous reviewers whose critical review helped improve the quality of the papers in this special issue. We would like to acknowledge the organizers and committee members of The Fourth International Workshop on Integrative Data Analysis in Systems Biology (IDASB 2013, in conjunction with IEEE Conference on Bioinformatics and Biomedicine, December 18–21, 2013, Shanghai, China) for their efforts to provide an international cross-disciplinary forum on systems biology, through which this special issue was made possible.

Huiru Zheng
Rui Jiang
Zhongming Zhao

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 369681, 16 pages
http://dx.doi.org/10.1155/2014/369681

Review Article

Advances and Computational Tools towards Predictable Design in Biological Engineering

Lorenzo Pasotti[1,2] and Susanna Zucca[1,2]

[1] *Laboratory of Bioinformatics, Mathematical Modelling and Synthetic Biology, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy*
[2] *Centre for Tissue Engineering, University of Pavia, 27100 Pavia, Italy*

Correspondence should be addressed to Lorenzo Pasotti; lorenzo.pasotti@unipv.it

Received 3 April 2014; Accepted 9 June 2014; Published 3 August 2014

Academic Editor: Huiru Zheng

The design process of complex systems in all the fields of engineering requires a set of quantitatively characterized components and a method to predict the output of systems composed by such elements. This strategy relies on the modularity of the used components or the prediction of their context-dependent behaviour, when parts functioning depends on the specific context. Mathematical models usually support the whole process by guiding the selection of parts and by predicting the output of interconnected systems. Such bottom-up design process cannot be trivially adopted for biological systems engineering, since parts function is hard to predict when components are reused in different contexts. This issue and the intrinsic complexity of living systems limit the capability of synthetic biologists to predict the quantitative behaviour of biological systems. The high potential of synthetic biology strongly depends on the capability of mastering this issue. This review discusses the predictability issues of basic biological parts (promoters, ribosome binding sites, coding sequences, transcriptional terminators, and plasmids) when used to engineer simple and complex gene expression systems in *Escherichia coli*. A comparison between bottom-up and trial-and-error approaches is performed for all the discussed elements and mathematical models supporting the prediction of parts behaviour are illustrated.

1. Background

In order to handle complexity in the design of customized systems, engineers usually rely on a bottom-up approach: components are quantitatively characterized and the output of an interconnected system is predicted from the knowledge of individual parts function [1]. This process is applied in all the fields of engineering and is useful to hide the complexity of the individual components functioning, thus using them as input-output modules [2].

This strategy is successful only in a modular framework, where parts behaviour does not change upon interconnections and, in general, when the same parts are reused in a different context [3, 4]. Even if this property does not persist, the bottom-up approach is still feasible when engineers are able to predict how parts behaviour varies as a function of environmental changes or interconnections [5]. In electronics, examples of the latter situation are resistors: they are characterized by an electrical resistance, which does not change upon connection in different circuits. However, it is well established that resistance changes as a function of temperature and, for this reason, datasheets of electric components report the temperature-resistance characteristic in order to make the output of complex circuits predictable when used in different environments. Another example is a circuit with a nonzero impedance; it can exhibit a different input-output behaviour when interconnected to different loads. However, it is still possible to predict the output of such interconnected systems since mathematical models of electrical circuits are able to describe voltage and current throughout the network.

Mathematical models are widely used in many areas of engineering to support the early design steps of a system, guide the debugging process, measure nonobservable parameters, and finally predict the quantitative behaviour of systems composed by precharacterized parts. Likewise,

models also play an important role in a biological systems framework; in fact, they are often used to study complex metabolic interactions, like those occurring in disease conditions to understand the underlying processes and/or predict the effect of drugs [25]. Some mathematical models of biological/physiological systems have also been approved by the US Food and Drug Administration (FDA) for use in simulated clinical trials, thus enabling researchers, for example, to support or even skip expensive *in vivo* trials [26].

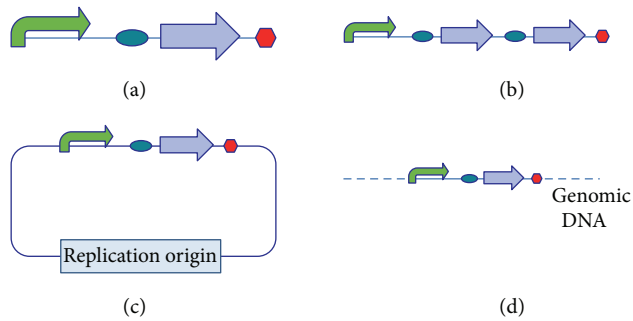
Synthetic biology aims to realize novel complex biological functions with the same principles on which engineering disciplines lay their foundations: modularity, abstraction, and predictability [2, 27, 28]. As a result, synthetic biologists so far have mainly focused on the definition of biological parts and on their abstraction and standardization, in order to deal with well-defined components with specific function [29]. This process has brought to the creation of biological parts repositories including DNA parts that can be shared by the scientific community, like the MIT Registry of Standard Biological Parts [30–32], to standardized and easy-to-automate DNA assembly strategies [33–35], and to standard measurement methodologies to share characterization results of parts, like promoters [36, 37]. Researchers have also focused on the realization of engineering-inspired functions to learn the complexity that could be reached in a biological context. Towards this goal, researchers built up devices that implement logic gates and functions [19, 38–41], memories [42], oscillators [43–45], other waveform generators [46, 47], signal processing devices [48–50], and the like. Many of them relied on mathematical models to support the early design steps and to capture the behaviour of the designed circuit. For example, two of the synthetic biology milestones are a genetic toggle switch [42] and an oscillator (the *repressilator*) [43], both implemented in *Escherichia coli* via genetic networks of properly connected transcriptional regulators. A semiquantitative investigation of the features required for a correct circuit behaviour was performed via mathematical models, by using dimensionless equations or reasonable parameter values. Thanks to the model analysis, the authors could learn useful guidelines for correct design of circuits exhibiting the desired functioning, for example, fast degradation rates of repressor proteins encoded in the oscillatory network [43].

The realization of complex functions has brought to some biological systems of high impact. An engineered pathway was implemented in recombinant yeast to produce the antimalarial drug precursor artemisinin [51]; a biosensor-encoding genetic device was implemented in microbes to detect arsenic in drinking water and to provide a colour change of its growth medium as visual output [52, 53]; microbes were recently engineered to produce bioethanol from algal biomass [54] or advanced fuels from different substrates [55].

However, despite many examples of complex engineering-inspired function implementation and also of industrially relevant solutions to global health, environmental, and energy problems, a rigorous bottom-up design process is not currently adopted because the predictability boundaries still have to be clearly defined [3, 56, 57]. The high potential of synthetic biology strongly depends on the achievement of such task [58]. Trial-and-error approaches represent an alternative: if synthetic biologists cannot design a system from the bottom-up, they can rely on random approaches, where, for example, circuit components are mutated and the best candidate implementing the function of interest is selected [38, 59, 60]. Depending on the reliability of predictions and of mathematical models, this process could be completely random or partially guided. In general, trial-and-error approaches are time- and resource-consuming, and are characterized by a low efficiency. However, recent advances in the construction of biological systems, for example, DNA and/or strain production via automated procedures, may provide a good alternative to the rational bottom-up approach, especially when accurate, automated, and possibly low-cost screening methods are available to rapidly evaluate the output of the constructed circuits [60].

This review discusses the predictability issues of basic biological parts (promoters, ribosome binding sites—RBSs, coding sequences, transcriptional terminators, and plasmids) when used to design the desired biological function in the form of a simple or complex gene expression system. Even though synthetic biological systems may be implemented in several organisms (or even *in vitro* [61]) and may have disparate architectures and regulatory mechanisms [62, 63], the review will focus on predictability of parts *in vivo* in the *E. coli* bacterium, according to the biological information flow described in the central dogma of molecular biology [64]: protein-coding DNA sequences (herein called genes) are transcribed into mRNA molecules, which are converted into proteins by ribosomes, and, finally, DNA sequences can be replicated in living cells to propagate the encoded function to the progeny. Thus, in the considered framework, the possible basic architectures are shown in Box 1: promoters can trigger the expression of a single gene (monocistronic architecture) or a set of genes (polycistronic or operon architecture), each gene is transcribed with its RBS upstream and finally terminators stop transcription. Ribosomes complete the process by translation of mRNA molecules into the proteins of interest, from the start codon (generally AUG) to the stop codon (generally UAA). Complex genetic circuits can be realized with a set of such gene expression units, implementing the interactions of interest and giving the desired product as output. Genetic circuits can be placed on a plasmid vector or otherwise they can be integrated into a target position of the bacterial chromosome.

Even if other classes of parts can be used to construct complex genetic systems and other elements can also affect circuit behaviour, we will focus only on the abovementioned genetic parts and architectures, given a specific strain and environment. Other important contexts, like the host (the reciprocal variation of parts behaviour and host metabolism when a circuit is incorporated), environmental (the reciprocal variation of parts behaviour and environmental parameters), ecological (changes of synthetic circuit and surrounding community parameters, as well as strains fitness), and evolutionary (changes of DNA composition) contexts are reviewed elsewhere [57]. Other reviews are



The simplest expression system(panel A) includes a promoter (curved arrow), a ribosome binding site (RBS, oval), a gene (straight arrow) and a transcriptional terminator (hexagon). A more complex architecture has an operon structure (panel B), where multiple genes, with their own RBSs, are expressed by a single promoter; a two-gene operon is shown. Individual or composite gene expression systems can be placed in a plasmid vector (panel C), which is maintained by the host strain at a specific copy number per cell, depending on the particular replication origin, or in the bacterial chromosome (panel D). Dashed line indicates the genomic DNA of the host.

Promoters control the transcription of the downstream DNA. Their sequence determines the information on transcriptional strength and regulation by other factors. Different promoter classes are present in E. coli, according to the specific sigma factor that determines their regulation. Other factors, such as activator/repressor proteins, which are widely used in synthetic circuits, can provide an additional degree of control.

RBSs are small sequences that are placed upstream of a gene. They are transcribed into mRNA but are not translated, since protein synthesis starts with the AUG codon of the coding sequence. RBS sequence determines the binding affinity between ribosome and mRNA, thus affecting the translation initiation rate. In particular, mRNA binding to ribosomes occurs at the Shine-Dalgarno region of an RBS with the rRNA of the 30S ribosomal subunit of E. coli.

Transcriptional terminators of the rho-independent class exploit hairpin structures to stop RNA polymerase activity during transcription. This class of terminators is commonly used in synthetic circuits, as opposed to rho-dependent terminators that mediate transcription stop via an ATP-dependent process, which has not been completely understood.

Genes encode the desired proteins, which can actuate disparate specific regulatory or metabolic functions. Gene sequence can affect the mRNA folding, which may result in different translation initiation rate and different decay rate of the transcript. Moreover, codon usage bias determines preferences of specific codons over others for a given organism. For these reasons, two identical proteins encoded by two different genes (designed exploiting genetic code redundancy), can have extremely diverse synthesis rates.

Box 1: Genetic parts and architecture of a gene expression system.

complementary to the present work, describing software tools for parts/pathway identification [65] and cellular behaviour modelling at different scales [65–67].

Each of the biological parts and architectures described in Box 1 will be considered. We will discuss to which extent their function can be predictable and then a comparison between bottom-up and trial-and-error approaches will be carried out. For each part and architecture, the contribution of mathematical models supporting the prediction of circuit behaviour will be highlighted. Even though many computer-aided design (CAD) tools are available for synthetic circuits [68], only mathematical analysis tools (also including tools from the field of systems biology) and predictive models of parts function will be considered, while no software tool for database access/development or for the assembly process support [65] will be taken into account. In particular, the considered tools can be ordinary differential equation (ODE) models (or derived steady-state equation models) based on empirical or mechanistic functions, or predictive models able to infer parts behaviour given their sequence and/or their DNA context.

2. Research Studies and Tools to Support Bottom-Up Design

The kit of parts, architectures, and contexts available to synthetic biologists will be discussed. Then, interconnection issues will be considered. A summary of the selected methods and tools available for parts/devices quantitative prediction is reported in Table 1.

2.1. Promoters. Promoters are intrinsically context-dependent parts, since it is known that their upstream and downstream elements may affect transcriptional activity [69–73]. The research studies on the predictability of promoters

Table 1: Selected computational methods and tools that support the bottom-up design in biological engineering.

Part, architecture or context	Description	Reference
Promoters	Strength prediction tool for sigmaE promoters, using a position weight matrix-based core promoter model and the length and frequency of A- and T-tracts of UP elements.	[6]
	Strength prediction tool for sigma70 promoters, using partial least squares regression.	[7]
Promoter-RBS pairs	Strength prediction tool for sigma70 promoter-RBS pairs, using an artificial neural network.	[8]
RBSs	RBS Calculator: a web-based tool for RBS strength prediction and forward engineering, frequently updated and able to design RBS libraries.	[9]
	RBS Designer: a stand-alone tool for RBS strength prediction and forward engineering, it considers long-range interactions within RNA and it can predict the translation efficiency of mRNAs that may potentially fold into more than one structure.	[10]
	UTR Designer: a web-based tool for RBS strength prediction and forward engineering, able to design RBS libraries and with the codon editing option to change RNA secondary structures.	[11]
Genes	GeMS: web-based tool for gene design, using a codon optimization strategy based on codon randomization via frequency tables.	[12]
	Optimizer: web-based tool for gene design using three possible codon optimization strategies: "one amino acid-one codon", randomization (called "guided random") and a hybrid method (called "customized one amino acid-one codon").	[13]
	Synthetic Gene Designer: web-based tool for gene design with expanded range of codon optimization methods: full ("one amino acid-one codon"), selective (rare codon replacement) and probabilistic (randomization-based) optimization.	[14]
	Gene Designer: stand-alone tool for gene design using a codon randomization method based on frequency tables and with the possibility to filter out secondary structures and Shine-Dalgarno internal motifs.	[15]
Terminators	Termination efficiency prediction tool based on a linear regression model using a set of sequence-specific features identified via stepwise regression.	[16]
	Termination efficiency prediction tool based on a biophysical model using a set of free energies, previously identified as important features.	[17]
Interconnected networks	A range of empirical or mechanistic ODE or steady-state models can be used to predict complex systems behaviour from the knowledge of individual parts/devices parameters.	[5, 18–21]
Architecture	Protein expression prediction for the first gene of an operon, given the downstream mRNA length, via a linear regression model.	[22]
Context	Mechanistic ODE models where the DNA copy number is explicitly represented.	[23]
	Protein expression prediction tool, based on linear regression model, given the chromosomal position of the gene and its orientation.	[24]

have focused on their context-dependent variability and on activity prediction given their nucleotide sequence. Context-dependent variability studies aim to evaluate whether promoters show the same activity in different contexts, for example, when promoters have different sequences upstream, when expressing different genes/mRNAs, or when other independent gene expression cassettes are present in the same circuit. Generally, the activity of a set of promoters can be indirectly measured via reporter proteins, provided that the downstream sequences are the same (i.e., identical RBSs, reporter genes, terminators, and similar transcription start sites—TSSs) so that mRNA primary and secondary structures do not significantly vary among the promoter measurement systems [37]. Using the same architecture, the activity can be evaluated via qPCR, by directly measuring the mRNA level [74]. Davis et al. [73] quantified a set of constitutive promoters and found that activity was affected up to 4-fold when a specific upstream (UP) sequence is placed before promoters, even though in some cases activity was not affected. Other studies showed that the upstream sequence-dependent activity change could be as high as 300-fold and the consensus sequences that can affect such different transcriptional activity were identified [72, 75, 76]; this effect was observed when using the rrnB P1 promoter, but activity change was also observed for the lac promoter. On the other hand, specific "anti" sequences downstream of promoters can limit the RNA polymerase escape process, thus affecting promoter activity [77]; such elements were found to decrease sigma70 and sigma32 promoter activity up to 10-fold [69]. Davis et al. also tested the effects of different sequences flanking promoters downstream, including an "anti" sequence or different reporter genes (GFP, dsRed, and Gemini) with the same RBS, yielding an activity change up to 2-fold [73]. A similar fold change was observed in analogous experiments, where Martin et al. [78] tested GFP, lacZ-alpha, and Gemini as reporter genes. In their work, however, the 2-fold difference persists for the strongest promoter, which might be affected by an excess of the lacZ-alpha fragment compared to the omega fragment needed for complementation. A study of our group [20] yielded a lower estimate of activity change for a set of 5 widely used promoters expressing the green fluorescent protein (GFP) with the BBa_B0032 RBS or the red fluorescent protein (RFP) with the BBa_B0032 or BBa_B0034 RBS: only one of the tested promoters showed a significant activity change among the three conditions, with a coefficient of variation (CV) of 22%. The abovementioned studies expressed promoter activities in RPUs, in order to provide comparable measurements among the different reporters used. Recent advances in DNA synthesis, assembly, and high-throughput characterization techniques enabled the quantification of very large libraries of single gene expression cassettes composed by different promoters, RBSs, and target genes, by measuring the fluorescence of reporter gene, as well as mRNA level via qPCR or next generation sequencing. In particular, Kosuri et al. performed the so far largest scale experimental study, where 114 promoters and 111 RBSs were combined upstream of a GFP gene [60]. Promoters were found to trigger consistent RNA levels of the downstream transcript among the different RBS-gene combinations. By using an ANOVA model for data interpretation, it was found that promoter sequence accounted for about 92% of total variability of mRNA level, demonstrating that promoters are the main factors affecting mRNA level, even though they expressed different mRNAs. RBSs accounted for 4% of total variability, which could be due to transcription rate modulation by the sequence downstream of promoter or to other phenomena not involving transcription, such as RBS-dependent mRNA degradation or sequestration (see discussion in Section 2.2.)

The majority of flanking sequence-dependent studies on promoters are relative to downstream sequences, while upstream sequences are less frequently studied. Even though highly stimulatory or inhibitory effects may be obtained via UP or "anti" sequences, promoters were found to change their activity within a reasonably low fold-change when not flanked by such difficult elements.

Although such data gave a significant contribution towards the understanding of promoter reusability, gene expression systems composed by independent expression cassettes are not similarly well studied and could yield unpredictable effects. Hajimorad et al. [79] studied the mRNA levels produced by different gene expression cassettes to test the superposition of the effects in synthetic biological systems at different copy number levels; they found conditions where even three cassettes could provide predictable levels of mRNA, while, in other configurations, cassettes could not be considered as modular systems. Similarly, our group [20] used two cassette-systems expressing GFP and RFP under the control of a set of promoters, detecting fluorescence as output. Cassette position was also studied. Context-dependent variability was higher than for individual cassette expressing different reporters (maximum CV of 33% versus 22%). A part of this variability could be explained by a different upstream sequence; that is, promoters could be flanked by the transcriptional terminator of the upstream cassette or by the plasmid sequence upstream of the cloning site.

Activity prediction studies given the nucleotide sequence of promoters have not yet produced accurate tools for the widely used sigma70 promoters. Promoter strength can be affected by many sequence features, which are not completely understood yet, including the −35/−10 sequences, the spacer between them and the above discussed flanking sequences. Recent efforts towards prediction include the works of Rhodius et al. [6, 80], who developed position weight matrix-based models to predict the activity of sigmaE promoters as a function of their sequence, as well as their flanking sequences (UP elements), with good predictive performance ($r = 0.86$ after cross-validation) [6]. However, the same methods are not likely to work for sigma70 promoters due to their complex structure [80]. De Mey et al. used partial least squares (PLS) regression to classify promoter strength as a function of nucleotide sequence [7]; this approach accurately predicted the activity of 6 out of 7 promoters used as a test set. Meng et al. developed an artificial neural network (ANN) to predict the strength of regulatory elements composed by a promoter and an RBS [8]; this approach brought to the accurate prediction of an initial test set of 10 promoter-RBS pairs ($r = 0.98$) and good performance was also obtained on

a second set of 16 newly constructed pairs. The described tools provided promising results but additional work is needed to independently validate such methods on other datasets and to fully understand promoter sequence features.

In summary, reproducible context-dependent variability studies should be performed to fully understand the factors affecting promoter activity in individual expression cassettes and in multiple cassette systems. Large libraries of parts are now affordable and, for this reason, the analysis of such factors will be facilitated, as well as activity prediction given promoter sequence. Standard [37] and multifaceted [74] characterization approaches have been proposed to provide robust measurements that can be shared and reproduced in many laboratories.

2.2. RBSs. RBSs are strongly context-dependent elements, since their surrounding sequences can affect ribosome binding and, as a result, the translation initiation rate per transcript. In particular, even a few nucleotide changes in the RBS or in the surrounding sequences can dramatically affect translation [10] and the use of different genes downstream of an RBS can provide completely different translational efficiencies [81]. Given the sequence of a gene and its 5′UTR, biophysical models have been used to predict the translation initiation rate by modelling local and global folding, as well as the interaction between RBS and 16S ribosomal RNA. Computational tools, such as the RBS Designer (stand-alone application, [10]), the RBS Calculator (web-based application, [9]), and the UTR Designer (web-based application, [11]) are available to perform such tasks. They take into account the 5′UTR sequence, as well as the first portion of coding sequence to predict the translation initiation rate level. The RBS Calculator and UTR Designer use similar biophysical thermodynamics-based models, while the RBS Designer uses a steady-state kinetic model of stepwise-occurring reactions [82, 83]. These tools showed similar and reasonably good predictive performance ($r^2 > 0.8$) and can also be used to forward-engineer novel RBSs with a desired strength [83]. They differentiate for the use of different external tools for energy computation [83] and for some specific peculiarities; for example, RBS Calculator provides indication of confidence and it is frequently updated [84], RBS Designer considers long-range interactions within RNA and can predict the translation efficiency of mRNAs that may potentially fold into more than one structure, while UTR Designer enables codon editing to minimize secondary structures [83]. Other efforts towards RBS prediction include an artificial neural network, already cited above, to evaluate the strength of promoter-RBS pairs [8].

The RBS Calculator is one of the most commonly used tools in the synthetic biology community: it was used in basic research studies to tune the response of a synthetic AND gate [9], to generate a set of RBSs of graded strengths to evaluate the transcription/translation processes [85], and to test DNA assembly platforms [33, 35], as well as in applied research to optimize biosynthetic pathways [86, 87]. Although it was proved to be useful to guide the choice of proper RBS sequences given a downstream gene, its accuracy is limited and additional tools should be developed to improve the predictability of RBSs [57, 81].

RBSs could also affect the mRNA decay rate by causing different secondary structures [16]. In addition, Kosuri et al. also observed a mutual interaction between transcription and translation: in fact, translation efficiency can affect mRNA levels, probably because the most translated mRNA molecules are protected from degradation, compared to the least translated mRNAs [60].

In summary, as in the case of promoters, large datasets have been useful to show the contributions of different context-dependent factors. Due to the strong context-dependent nature of RBSs, experimental studies mainly focused on flanking sequences, while the evaluation of RBS modularity in complex circuits still needs to be studied.

2.3. Genes. Given a target protein, its coding sequence can affect both transcription and translation processes [15, 88]. As described above, mRNA secondary structures could affect mRNA degradation and limit RBS accessibility to ribosomes and, in addition, AT-rich sequences can cause premature transcriptional termination [89]. Codon usage has been reported to affect the translation process [90]. In this framework, most of the efforts towards the prediction of the contribution of gene sequence to transcription/translation processes have focused on the development of gene optimization algorithms. To define them, several sequences need to be constructed to cover a sufficient number of hypotheses; although the cost of synthetic genes is greatly decreasing, gene synthesis still brings to expensive studies [88]. For this reason, the process of sequence optimization is not fully understood and no consensus rules have been found for gene optimization. Some research studies identified strong secondary structures as the primary limiting factors in protein synthesis [91], while other studies did not find a correlation between predicted secondary structure and expression level [92]. On the other hand, in some studies expression level has been found to correlate with the codon adaptation index (CAI) [93, 94], often used to express the codon bias of a gene towards common codons [95], while in other studies this correlation was null [88, 91]. The codon randomization method, where codons are extracted from codon usage frequency tables, was found to be superior to the "one amino acid-one codon" strategy, where the CAI is maximized [15, 92]. Finally, codon context, that is, the influence of codon pair usage, was found to affect protein expression, although no ready-to-use software tool is available to carry out an optimization procedure based on such feature [90].

All the features described above might be gene and variant dependent [88] and, for this reason, several studies should be conducted to identify the correct features of gene sequence affecting transcription, translation, and other processes. In particular, the simultaneous measurement of mRNA and protein level can provide exhaustive data to decouple the effects of gene sequence changes on cellular processes. In a large-scale study, performed by Goodman et al., a library of >14,000 expression systems was constructed to test the contribution of the N-terminal codons on gene

expression [96]; they measured DNA, RNA, and protein levels and confirmed that mRNA secondary structure is a crucial factor which can tune gene expression up to ~14-fold.

The research efforts carried out so far have brought to different gene optimization tools, currently used by synthetic biologists and gene synthesis companies to optimize protein expression, according to codon usage frequency tables, global GC content, minimization of hairpin structures within the gene, and/or of secondary structures in the N-terminal codons [97, 98]. The free software tools proposed in literature include, for instance, GeMS (web-based application, [12]), Optimizer (web-based application, [13]), Synthetic Gene Designer (web-based application, [14]), and Gene Designer (stand-alone application, [15]). All the tools mainly differentiate for their available options for designing genes (e.g., avoid unwanted restriction sites and inverted repeats, design framework of oligonucleotides for gene synthesis) and for their codon optimization strategy (e.g., "one amino acid-one codon" method, probabilistic methods, or hybrid solutions, based on codon frequency tables from different sources) to take into account codon usage and constraints. Because many available tools are proprietary of gene synthesis companies, an accurate comparison of the implemented methodologies is not feasible and, in addition, their performances still need to be experimentally evaluated on different gene sets.

In summary, although prediction tools have been proposed, no widely accepted algorithm is available to predict the effects of gene sequence on transcription, translation, or mRNA degradation.

2.4. Terminators. Rho-independent terminators are herein considered. Although very efficient terminators are available (e.g., the popular BBa_B0015 double terminator from the MIT Registry of Standard Biological Parts), the repeated use of a small set of elements in a genetic circuit may result in poor evolutionary stability [99, 100]. For this reason, reliable methods to design new terminators with predictable strength and methods to predict the efficiency of already existing terminators given their sequence are required.

Terminator efficiency can be characterized via an operon-structured measurement system, where a promoter drives the expression of two different reporter genes with the terminator sequence to be measured that is assembled between these two genes. The two reporter proteins are quantified and termination efficiency is computed from their values, considering the operon without the terminator of interest as a control [16, 17, 101].

Like promoters and RBSs, also terminator efficiency has been found to be dependent on the surrounding context. In particular, Cambray et al. [16] tested different minimal terminators, including only the hairpin and U-tail sequences and compared their termination efficiency to the respective full-length terminators. Efficiencies significantly changed between the two contexts for almost all the 11 tested terminators, demonstrating that sequences flanking the essential terminator parts are crucial. The authors also used a multiple linear regression model to build up a predictive tool for transcriptional termination given the terminator sequence, using a set of features identified via stepwise regression, but the resulting predictor gave poor performance on the 54 terminators used ($r = 0.61$ after cross-validation). Only by excluding the low efficiency terminators, low predicted folding frequency terminators, and extended terminators classes, the Pearson correlation coefficient r increased to 0.85 after cross-validation. Through a complementary approach, Chen et al. [17] experimentally characterized a large set of terminators (582) and analyzed how sequence features contribute to their strength. The dominant features were used to build up a biophysical model that aimed to capture termination strength (Ts) as a function of the U-tract, hairpin loop, stem base, and A-tract-free energies. The model was used to fit via linear regression the experimentally determined Ts, yielding a squared r value of 0.4, which results in low predictive performances. Although not currently available to users, the tools developed in the above publications [16, 17] can be implemented through the provided regression coefficients, web-based nucleic acid folding tools, and specific indexes computed from terminator sequences. These two recent studies relied on experimental measurements performed via the abovementioned operon structure with reporter genes. However, Cambray et al. constructed measurement plasmids with RNAse sites flanking the terminator to be measured, in order to avoid terminator-dependent mRNA folding, which might affect the translation efficiencies of the two reporter genes. The authors tested RFP-GFP and GFP-RFP operons with terminators flanked by RNAse III, RNAse E, or non-functional RNAse III sites. The configuration giving the lower coefficient of variance for the upstream gene level was the RFP-GFP operon with RNAse III sites, which was used for all the characterization experiments of their paper. Conversely, Chen et al. used a GFP-RFP operon without RNAse sites, since they found that, in their configuration, RNAse E sites presence affected the downstream gene expression. In light of these findings, a standard measurement method for terminators still needs to be defined in order to enable reliable quantifications and to avoid potential mechanisms that may complicate the measurement of terminator efficiency, for example, promoters that might arise at the interface of the terminator to be measured and the downstream gene of the operon [17].

In summary, sequence features affecting terminators behaviour have been recently evaluated on large datasets, but predictive models with good performances are not available yet, demonstrating that different models and additional knowledge on transcriptional termination are needed, as well as a standardized setup for experimental measurements.

2.5. Interconnected Networks and Retroactivity. In the philosophy of bottom-up composition of biological systems, arbitrarily complex networks are considered as black-box modules that can be interconnected. Their characterization can provide the essential elements to describe their steady-state and dynamic behaviour. In a modular framework, such knowledge enables the prediction of composite networks functioning. To quantitatively test the modularity boundaries of biological systems, recent studies have focused on the

characterization of systems subparts and on the prediction of the behaviour of composite systems, obtained upon their interconnection. Wang et al. [18] tested different regulated promoters (inducible by arabinose, AHL, and IPTG) as the inputs of AND/NAND gates, whose output was visualized via GFP at two different temperatures. After a fitting process involving one specific configuration (i.e., one of the cited input modules), the fluorescence output of the other configurations was predicted from the individual characterization of input devices and AND/NAND gates. Experimental data and predictions exhibited a Pearson correlation coefficient of 0.86 to 0.98, even though some specific input combinations yielded highly different values. Moon et al. [19] constructed and characterized a set of AND gates. Then, they used them to engineer composite two layered logic functions: a 3-input system including 3 input devices connected to two AND gates and a 4-input system including 4 input devices and 3 AND gates. The latter represented one of the largest genetic programs built up so far, with a total of 11 regulatory proteins, 21 kbp-length on three plasmids. The basic AND gates were individually characterized as before and the output of the complex 3- and 4-input systems was predicted and compared with experimental data. The 3-input system yielded a lower deviation between prediction and data, compared to the 4-input system. Our group also faced prediction problems with simple interconnected networks composed by an input device (inducible promoters or constitutive promoters of different strengths) assembled with a TetR-based NOT gate which provides GFP as output [20]. The individual input devices were characterized via RFP measurements and the steady-state transfer function output of the NOT gate driven by each of the input systems was quantified. These data were fitted with a Hill function: they had similar maximum activity and Hill coefficients, while the switch point varied about 44%, which was considered as an estimate of interconnection error with these elements.

The mentioned studies evaluated interconnection-dependent variability in considerably complex systems but they did not characterize the causes of such deviations. One of the best characterized and formalized interconnection errors is retroactivity, a phenomenon that extends the electronic engineering notion of impedance or loading to biological systems [5]. The functioning of a given system can change when a downstream or upstream system is connected, for example, because of unwanted sequestration of transcription factors by the connected modules. In this case, the individual systems cannot be considered to be modular; however, given the knowledge of the parts to be combined, such unwanted interactions can be modelled, thus having an interconnected system with predictable behaviour. Jayanthi et al. [21] experimentally tested a model system including an ATc-inducible LacI production module connected to a lac-repressible promoter with GFP downstream. This composite system was placed in a medium-copy plasmid and tested individually or in presence of a downstream "client," including lac operator sites in a high-copy plasmid, thus providing additional binding sites for LacI. The presence of the client significantly affected the induction and deinduction dynamics. This phenomenon was captured by a mechanistic model describing the LacI-occupied DNA sites upstream of GFP and in the client binding, as a function of ATc induction.

2.6. Circuit Architecture. Most of the research studies described above are based on single gene cassettes. The polycistronic operon structure could be preferred when expressing genes carrying out similar functions that can be controlled by the same promoter. Although predictable RBS tuning in operons has been reported [87], the prediction of protein levels encoded by genes in operons is not trivial and cannot be simply inferred by the protein levels of individual gene cassettes. In particular, the specific operon structure can affect mRNA degradation rate and ribosome accessibility. Lim et al. developed and experimentally tested a mathematical model of transcription and translation coupling, which predicts the protein level encoded by the first gene as a function of the operon length [22]. They found and predicted protein level variations up to 2- to 3-fold. In a complementary framework, Levin-Karp et al. studied the translational coupling of an operon, that is, the mutual relationships between the translation efficiencies of neighbouring genes [102]. They individuated a >10-fold change for the protein level encoded by the second gene as a function of the translation rate of the first gene. However, the findings of Lim et al. and Levin-Karp et al. were not valid for all combinations of genes and the same phenomena were not observed in different studies [61, 102].

The measurement of mRNA levels of a transcribed operon has been useful to decouple the effects of RNA stability and translation rate change [102]. In summary, other mathematical analyses are needed to develop predictive tools that can guide biological engineers in the composition of operon structures with quantitatively predictable function, which can be inferred by the knowledge of promoter, RBSs, gene sequence, genes position, operon length, and other possible features [22].

2.7. Genetic Context. The context in which a gene expression cassette or a complex circuit is placed can affect its quantitative behaviour. Genetic contexts include plasmids replicating at different copy numbers per cell or the bacterial chromosome. Given a single gene expression cassette, plasmid sequence can affect promoter or terminator activity by means of the sequences flanking the cloning site, as described above for these two part classes. Moreover, intuitively, DNA copy number determines different levels of all the species (mRNA and protein), but such levels could be unpredictable, since cells may exhibit metabolic overloading when copy number is increased, thus showing nonlinear changes. This effect is commonly observed in expression cassettes at high copy number [20, 79, 103] and needs to be characterized when the cassette copy number is to be tuned. Furthermore, plasmid copy number can be intrinsically noisy [104, 105] and can also change when multiple plasmids are incorporated in the same cell [106]. To test the latter case, Lee et al. [106] showed that low copy plasmids with the heat-sensitive pSC101 replication origin maintain their copy number (about 5 copies per cell) in single plasmid systems and in 3-plasmid systems, while plasmids with the medium or high copy replication origins

(p15A and ColE1, resp.) showed copy number increase when used in the 3-plasmid system compared to the single plasmid system.

Mathematical models of gene regulatory networks often use empirical Hill functions to describe activation or repression of cellular species, but DNA copy number is not explicitly present in the equations [23, 103]. For this reason, even by assuming a linear change of cellular species as a function of DNA copy number, mechanistic mathematical models should be defined to easily study the copy number effects. Although such models are also widely used to describe biochemical reactions, they are more difficult to study and identify than empirical models, thus requiring additional work to fully characterize the system of interest. Mileyko et al. used such class of models to study the copy number effects on different gene network motifs [23].

The integration of the desired expression cassette in the bacterial chromosome determines the maintenance of its DNA in a single copy, replicated with the genome. However, the quantitative behaviour of parts in the genomic context can be difficult to predict. For example, the real copy number of the desired DNA could change when integrated in different genomic positions because the sequences near the bacterial replication origin are expected to be replicated earlier than the other sequences [24, 107] and thus the specific DNA segment is actually present in the cell at a slightly higher copy number, on average. The complexity of genomic context is not limited to this effect and other not fully understood phenomena could limit the prediction of an integrated cassette. For example, transcriptional read-through from flanking genomic cassettes could affect the expression of the synthetic cassette.

3. Trial-and-Error Approaches

The design of a desired biological function can be achieved by randomly changing its DNA-encoded elements. In particular, promoters, RBSs, architectures, and contexts are varied, via disparate experimental methods, and the resulting circuit is screened. The success of all these methods relies on parts generation and screening efficiency, which should allow an easy and high-throughput construction and recognition of the desired phenotype [60]. Here, only representative studies are illustrated, which randomly optimize promoters, RBSs, genes, architectures, and context towards a target circuit/pathway functioning.

Promoters upstream of one or more target genes is randomly changed by directly synthesizing new promoter sequences or by assembling the genes under the control of a collection of promoters. In the first case, degenerate primers can be used to insert a new random promoter sequence upstream of a gene [108]. In the second case, promoters from existing collections of parts [55] or random fragments [109, 110] can be used in the same manner and the resulting constructs are screened. In this latter case, the characterization of promoters (or the quantification of the transcriptional activity of random fragments) is not required, because only the circuit outcome is considered to optimize the process. These two methods can be combined by producing libraries of synthetic random promoters, when required with the desired design constraints (e.g., the desired operator sites) [74, 111], that are screened by reporter genes to yield a collection of parts with diverse and graded activity; then, elements can be randomly assembled to tune the desired circuit/pathway [74, 111]. Such procedure could be partially rational: inducible promoters can be used to probe the optimal activity of a target gene and only a subset of the candidate newly generated promoters, having a constitutive activity similar to the optimal one, can be tested [20, 112, 113].

By following a similar procedure, RBSs can be randomly changed and selected. Anderson et al. [38] and Kelly [101] repaired a nonfunctional AND gate and a logic inverter, respectively, by random mutagenesis of the RBS upstream of a regulatory gene. The two gates were nonfunctional because their activity range in input did not match the activity range provided by the upstream promoter used in the final interconnected circuit. The RBS sequence mutagenesis and screening process produced circuits with the expected behaviour. The use of existing collections of RBSs can also be exploited instead of creating new ones [42, 114]. The random mutagenesis of promoters and RBSs can be performed via different widely used molecular biology methods, including error-prone PCR or DNA amplification with degenerate primers. High-throughput techniques have been recently proposed to simultaneously mutate the sequence of several elements, also in the genome, via automated procedures. The multiplex automated genome engineering (MAGE) approach was used, coupled with a microfluidic automatic system and with degenerate single-stranded DNAs to enable the lycopene pathway optimization through RBS mutagenesis for 24 target genes in plasmid or genome [115].

Genes have been randomly mutated mainly to obtain different functional protein variants with improved performance [59]. Since this approach causes amino acid variation, instead of synonymous codon replacement, the resulting protein is different. Such approaches are beyond the focus of this review. Codon change studies, without affecting protein sequence, are not widely used and they are limited to the experimental works carried out to find gene optimization rules, as described in Section 2.3 of this review. Similarly, terminators are not commonly targeted for random mutations.

When dealing with polycistronic designs, the architecture of gene expression cassettes can be randomly varied by changing the position of the genes in an operon or by flanking genes with libraries of tunable intergenic regions (TIGRs) [116]. Since the target protein level produced by genes in operons is not currently predictable, the first, intuitive, method relies on random change of gene position. This, in several studies, yielded highly diverse protein levels among the shuffled constructs. For example, bicistronic operons including the 1a-hydroxylase, adrenodoxin, and NADPH-adrenodoxin reductase genes (called ADX and ADR), used as redox partners to characterize the 25-hydroxyvitamin D3 1a-hydroxylase gene, were switched (yielding ADX-ADR and ADR-ADX constructs) and both ADR and ADX expression levels varied up to 5-fold [117]. On the other hand, the use of TIGRs relies on the assembly of various control

elements (mRNA secondary structures, RNAse cleavage sites, RBS sequestering sequences, etc.) within operon genes. This random approach has proved to enable a >100-fold range of enzyme levels and a 7-fold improvement of productivity for a synthetic mevalonate pathway [116].

The genetic context can also be randomly optimized. Plasmid copy number change is an intuitive method to tune the output of circuits and pathway. Kittleson et al. [118] constructed different-allele (DIAL) strains that had the same genetic background except for an expression cassette providing different protein levels of a trans-acting replication factor (Pi or RepA); plasmids with the R6 K and ColE2 replication origins can be maintained at disparate copy number per cell levels, due to the regulation by Pi and RepA, respectively. The resulting strains were successfully used to optimize a violacein biosynthetic pathway. Considering genetic context at genomic level, different methods were used to optimize integration position and copy number of synthetic DNA-encoded production pathways via random approaches. Santos et al. developed a recombinase-assisted genome engineering (RAGE) approach, where lox sites, recognized by the Cre recombinase, are exploited to integrate very large synthetic DNA fragments into the desired genomic position, thus enabling the trial-and-error search among several predefined candidate loci [119]. They used it to optimize a 34 Kb heterologous pathway for alginate metabolism. On the other hand, the random insertion of the desired DNA parts is often carried out through transposable elements. By randomly optimizing promoter activity and genomic position at the same time, Yomano et al. optimized the expression of an ethanol production pathway [120]. In particular, they integrated a promoter-less 3-cistron ethanol production cassette in random positions of the strain of interest via a mini-Tn5 cassette (transpososome), relying on the random placement of the cassette under the control of promoters with optimal strength in the optimal genomic position.

Chromosomally integrated circuits or pathways can be also optimized by randomly changing their copy number. Methods to carry out this task rely on genomic integration of the DNA of interest together with an antibiotic resistance cassette; subsequently, recombinant strains are evolved in presence of increasing antibiotic concentration, to promote the tandem duplication of the recombinant DNA cassette, until a target efficiency is reached. This method has provided recombinant strains containing more than 25 copies of the DNA-encoded ethanol production pathway to be optimized [121, 122]. A further refinement of the methods was carried out by Tyo et al., where the chemically inducible chromosomal evolution (CIChE) was described [123]. It is analogous to the previously described procedure, but when the desired efficiency is reached the recA gene (promoting homologous recombination) is knocked out. CIChE was applied to poly-3-hydroxybutyrate (PHB) and lycopene production, yielding significant pathway improvement (4-fold and 60%, resp.). This method produced approximately 40 consecutive copies of the DNA-encoded pathway and 10-fold improvement on genetic stability [123].

4. Interventions on Circuit Structure to Improve Predictability

Although individual parts, networks, architectures, and contexts have the abovementioned predictability issues, several efforts have been undertaken to modify some of these elements to decrease their context-dependent variability and improve their predictability.

Davis et al. designed a set of insulated promoters that extend from −105 to +55 from the transcription start site [73]. These elements had a more predictable activity than noninsulated promoters when tested in different contexts. Mutalik et al. proposed a bicistronic design (BCD) of gene expression cassettes to effectively predict the translation initiation rate of a downstream gene [81]. This design includes a small open reading frame (ORF), with its own RBS, assembled downstream of the promoter of interest. The stop codon of this ORF is fused to the start codon of the gene of interest (thus having TAATG), which is assembled downstream. The RBS of the gene of interest is included in the small ORF upstream. With this design, inhibitory RNA structures around the gene of interest start codon or RBS are eliminated by the intrinsic helicase activity of ribosomes arriving at the stop codon of the upstream ORF. By forward-engineering an expression cassette via BCD, users should obtain the expected relative expression within 2-fold of the target value with 93% probability, which represents a great improvement over state-of-the-art predictive tools for RBSs [9, 81].

Qi et al. proposed the use of bacterial clustered regularly interspaced short palindromic repeat (CRISPR) pathway elements to engineer specific posttranscriptional cleavage of multigene operons to yield predictable expression of the individual genes, also when placed in different positions [124]. Via a complementary approach, Lou et al. used ribozymes, assembled downstream of a promoter, to improve the predictability of gene expression [125]; ribozymes cleave the mRNA eliminating their 5′ end and also act as transcription insulators.

Del Vecchio et al. [5, 126] proposed a system able to overcome retroactivity issues upon interconnection of biological systems, thus implementing a buffer (or insulator) device. It strongly relies on engineering-inspired insulators, such as noninverting operational amplifiers. The biological implementation of this mechanism includes phosphorylation-dephosphorylation reactions, which act with fast timescales, but it needs to be experimentally validated.

5. Conclusions

This review has described several aspects of the design of genetic circuits with predictable function. Bottom-up approaches have been recently investigated to mimic the traditional design processes in engineering areas. In this context, research studies have been carried out to evaluate the predictability boundaries of biological systems composed by precharacterized parts, providing the expected interconnection error, estimated from the study of model systems, and highlighting situations where circuits cannot behave

(i) What is still needed to fully understand the context-dependent variability of biological parts?
(ii) How can we develop accurate computational tools for the bottom-up biological engineering?
(iii) How can we support the construction of novel synthetic biology-based solutions?
(iv) How can we understand, model and predict cell-to-cell variability in the functioning of biological systems?
(v) If the fully predictable engineering of biological systems is not possible, how can we improve the efficiency of trial-and-error approaches?

Box 2: Outstanding questions.

as intended. Mathematical models support the bottom-up design steps, from the early feasibility study of complex functions to the quantitative prediction of circuit behaviour from the knowledge of basic parts function and, finally, to the debugging step.

To exploit the full potential of synthetic biology via an engineering-inspired bottom-up design of circuits, several challenges need to be faced. The main crucial issues identified in the context of this work are delineated in Box 2 in the form of outstanding questions and they are herein discussed.

Predictable biological engineering requires deepening our knowledge on context dependency and reusability of biological parts, by discovering the features that play important roles in parts function predictability. Technology advances in the DNA synthesis field can support the testing of large number of hypotheses by providing huge libraries of constructs at affordable price. In fact, although large-scale studies have been reported to support the investigation of different aspects of parts predictability [60, 81, 96], the cost and scale of DNA synthesis are still a major bottleneck for basic research, since many studies require a very large number of construct variants, as in the case of codon usage dependency in protein expression [88]. The development of high-throughput methods for parts measurement plays a complementary role, because multifaceted characterization of parts performance needs to be carried out. In particular, to fully characterize the activity of parts, the simultaneous quantification of DNA, RNA, and proteins is required to accurately decouple effects due to circuit copy number, transcription, and translation, to improve the knowledge of all the atomic steps involved in parts function. In addition, ad hoc experimental designs, data analysis tools, and mathematical models can support the above procedures; for example, models can be of help in the estimation of nonobservable parameters, useful to characterize parts function [36].

Empirical mathematical models of gene regulatory networks are currently used to summarize the function of parts and predict the quantitative behaviour of higher-order devices. Although they are widely used, in some cases mechanistic models could be more appropriate tools, such as in the study of DNA copy number variations or retroactivity effects. Other tools enable the prediction of parts activity from the knowledge of their nucleotide sequence. Although promising results have been obtained, particularly in the case of RBSs that are already optimized via these computational methods, these tools need to be significantly improved. The data and knowledge gained in the above "discovery" step are to be exploited in the development of predictive computational tools with greater accuracy than the current ones. In this context, novel tools can be based on the acquired biological knowledge, which will be used to define essential rules for parts function prediction or can be data-based, where machine learning methods are used to learn the relationships of interest for parts prediction. Context-dependent activity change of individual parts and mathematical models of interconnected networks should ultimately be integrated to contribute unique tools for interconnected circuit design from parts sequence.

In addition to existing parts prediction, an ambitious goal of synthetic biology is the construction of unnatural parts with finely tuned customized function. To this aim, the computational design tools need to be expanded to support the forward engineering of new components according to specific design rules, learned from data examples or from the acquired biological knowledge. Again, the currently available RBS design tools already enable the design of RBSs with desired strength, given the downstream gene sequence, although their performance needs to be significantly improved [57]. Specifically, the RBS Calculator computes novel RBS sequences with about 47% chance to show the target strength within 2-fold [81].

Even though most of our current biological knowledge is based on population-averaged data and central tendency values, cell-to-cell variability is a crucial issue and can bring to unpredictable system behaviour. Although the main aspects of this point are described elsewhere [127] and are beyond the scope of this review, we want to highlight that biological noise can be detrimental for circuits function, even when central tendency values are predictable. For this reason, the full characterization of biological components should also take into account cell-to-cell variability, which needs to be propagated throughout an interconnected network of well-characterized modules to obtain reliable quantitative predictions of network output.

In this review, trial-and-error approaches involving the random-based optimization of parts/circuit function have also been briefly illustrated. These approaches rely on

affordable parts construction methods and efficient high-throughput-compatible screening methods to select the best combination of genetic parts, while these approaches cannot be efficiently applied when this condition does not persist. The technology advances mentioned above could greatly support the generation of large libraries to be screened via appropriate high-throughput measurement techniques, even without significant improvements in biological discoveries about context-dependent variability. However, while the learning of predictability boundaries is expected to contribute definitive predictive tools to handle the complexity of biological systems, trial-and-error approaches do not ensure the success of synthetic biology. In fact, large numbers of candidate constructs can be built up, but high-throughput measurement methods are not always available for the quantitative evaluation of circuit activity and the impact of pure trial-and-error approaches remains limited to specific projects. For this reason, bottom-up approaches urgently need to be refined to exploit the full potential of synthetic biology. A mixture of prediction tools, even with nonoptimal accuracy, and trial-and-error approaches could rapidly boost the efficiency of biological engineering, by providing a smaller search space than fully random-based approaches.

Finally, intense interventions on genetic circuits have been reported, which can provide considerable improvements to the predictability of promoters, RBSs, architecture, and retroactivity issues in different contexts. Since such improvements are highly promising, these modifications should be used in different studies to demonstrate their benefits on large scale and they should be considered in all the previously mentioned issues.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. M. Sauro, "Modularity defined," *Molecular Systems Biology*, vol. 4, p. 166, 2008.

[2] D. Endy, "Foundations for engineering biology," *Nature*, vol. 438, no. 7067, pp. 449–453, 2005.

[3] R. Kwok, "Five hard truths for synthetic biology," *Nature*, vol. 463, no. 7279, pp. 288–290, 2010.

[4] M. Muers, "Synthetic biology: quality and quantity," *Nature Reviews Genetics*, vol. 14, no. 5, article 303, 2013.

[5] D. Del Vecchio, A. J. Ninfa, and E. D. Sontag, "Modular cell biology: retroactivity and insulation," *Molecular Systems Biology*, vol. 4, article 161, 2008.

[6] V. A. Rhodius, V. K. Mutalik, and C. A. Gross, "Predicting the strength of UP-elements and full-length *E. coli* σ^E promoters," *Nucleic Acids Research*, vol. 40, no. 7, pp. 2907–2924, 2012.

[7] M. De Mey, J. Maertens, G. J. Lequeux, W. K. Soetaert, and E. J. Vandamme, "Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering," *BMC Biotechnology*, vol. 7, article 34, 2007.

[8] H. Meng, J. Wang, Z. Xiong, F. Xu, G. Zhao, and Y. Wang, "Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network," *PLoS ONE*, vol. 8, no. 4, Article ID e60288, 2013.

[9] H. M. Salis, E. A. Mirsky, and C. A. Voigt, "Automated design of synthetic ribosome binding sites to control protein expression," *Nature Biotechnology*, vol. 27, no. 10, pp. 946–950, 2009.

[10] D. Na and D. Lee, "RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression," *Bioinformatics*, vol. 26, no. 20, pp. 2633–2634, 2010.

[11] S. W. Seo, J.-S. Yang, I. Kim, B. E. Min, S. Kim, and G. Y. Jung, "Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency," *Metabolic Engineering*, vol. 15, no. 1, pp. 67–74, 2013.

[12] S. Jayaraj, R. Reid, and D. V. Santi, "GeMS: an advanced software package for designing synthetic genes," *Nucleic Acids Research*, vol. 33, no. 9, pp. 3011–3016, 2005.

[13] P. Puigbò, E. Guzmán, A. Romeu, and S. Garcia-Vallvé, "OPTIMIZER: a web server for optimizing the codon usage of DNA sequences," *Nucleic Acids Research*, vol. 35, no. 2, pp. W126–W131, 2007.

[14] G. Wu, N. Bashir-Bello, and S. J. Freeland, "The Synthetic Gene Designer: a flexible web platform to explore sequence manipulation for heterologous expression," *Protein Expression and Purification*, vol. 47, no. 2, pp. 441–445, 2006.

[15] A. Villalobos, J. E. Ness, C. Gustafsson, J. Minshull, and S. Govindarajan, "Gene Designer: a synthetic biology tool for constructuring artificial DNA segments," *BMC Bioinformatics*, vol. 7, article 285, 2006.

[16] G. Cambray, J. C. Guimaraes, V. K. Mutalik et al., "Measurement and modeling of intrinsic transcription terminators," *Nucleic Acids Research*, vol. 41, no. 9, pp. 5139–5148, 2013.

[17] Y. J. Chen, P. Liu, A. A. K. Nielsen et al., "Characterization of 582 natural and synthetic terminators and quantification of their design constraints," *Nature Methods*, vol. 10, no. 7, pp. 659–664, 2013.

[18] B. Wang, R. I. Kitney, N. Joly, and M. Buck, "Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology," *Nature Communications*, vol. 2, no. 1, article 508, 2011.

[19] T. S. Moon, C. Lou, A. Tamsir, B. C. Stanton, and C. A. Voigt, "Genetic programs constructed from layered logic gates in single cells," *Nature*, vol. 491, no. 7423, pp. 249–253, 2012.

[20] L. Pasotti, N. Politi, S. Zucca, M. G. Cusella De Angelis, and P. Magni, "Bottom-up engineering of biological systems through standard bricks: a modularity study on basic parts and devices," *PLoS ONE*, vol. 7, no. 7, Article ID e39407, 2012.

[21] S. Jayanthi, K. S. Nilgiriwala, and D. Del Vecchio, "Retroactivity controls the temporal dynamics of gene transcription," *ACS Synthetic Biology*, vol. 2, no. 8, pp. 431–441, 2013.

[22] H. N. Lim, Y. Lee, and R. Hussein, "Fundamental relationship between operon organization and gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 26, pp. 10626–10631, 2011.

[23] Y. Mileyko, R. I. Joh, and J. S. Weitz, "Small-scale copy number variation and large-scale changes in gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 43, pp. 16659–16664, 2008.

[24] D. H. S. Block, R. Hussein, L. W. Liang, and H. N. Lim, "Regulatory consequences of gene translocation in bacteria," *Nucleic Acids Research*, vol. 40, no. 18, pp. 8979–8992, 2012.

[25] M. Simeoni, G. De Nicolao, P. Magni, M. Rocchetti, and I. Poggesi, "Modeling of human tumor xenografts and dose rationale in oncology," *Drug Discovery Today: Technologies*, vol. 10, no. 3, pp. e365–e372, 2013.

[26] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, "*In silico* preclinical trials: a proof of concept in closed-loop control of type 1 diabetes," *Journal of Diabetes Science and Technology*, vol. 3, no. 1, pp. 44–55, 2009.

[27] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss, "Synthetic biology: new engineering rules for an emerging discipline," *Molecular Systems Biology*, vol. 2, 2006.

[28] D. E. Cameron, C. J. Bashor, and J. J. Collins, "A brief history of synthetic biology," *Nature Reviews Microbiology*, vol. 12, pp. 381–390, 2014.

[29] G. M. Church, M. B. Elowitz, C. D. Smolke, C. A. Voigt, and R. Weiss, "Realizing the potential of synthetic biology," *Nature Reviews Molecular Cell Biology*, vol. 15, pp. 289–294, 2014.

[30] MIT, Registry of Standard Biological Parts, http://partsregistry.org/.

[31] R. P. Shetty, D. Endy, and T. F. Knight, "Engineering BioBrick vectors from BioBrick parts," *Journal of Biological Engineering*, vol. 2, article 5, 2008.

[32] J. C. Anderson, J. E. Dueber, M. Leguia, G. C. Wu, A. P. Arkin, and J. D. Keasling, "BglBricks: a flexible standard for biological part assembly," *Journal of Biological Engineering*, vol. 4, article 1, 2010.

[33] J. E. Norville, R. Derda, S. Gupta et al., "Introduction of customized inserts for streamlined assembly and optimization of BioBrick synthetic genetic circuits," *Journal of Biological Engineering*, vol. 4, article no. 17, 2010.

[34] M. A. Speer and T. L. Richard, "Amplified insert assembly: an optimized approach to standard assembly of BioBrick genetic circuits," *Journal of Biological Engineering*, vol. 5, article 17, 2011.

[35] M. Leguia, J. A. N. Brophy, D. Densmore, A. Asante, and J. C. Anderson, "2ab assembly: a methodology for automatable, high-throughput assembly of standard biological parts," *Journal of Biological Engineering*, vol. 7, no. 1, article 2, 2013.

[36] B. Canton, A. Labno, and D. Endy, "Refinement and standardization of synthetic biological parts and devices," *Nature Biotechnology*, vol. 26, no. 7, pp. 787–793, 2008.

[37] J. R. Kelly, A. J. Rubin, J. H. Davis et al., "Measuring the activity of BioBrick promoters using an in vivo reference standard," *Journal of Biological Engineering*, vol. 3, article 4, 2009.

[38] J. C. Anderson, C. A. Voigt, and A. P. Arkin, "Environmental signal integration by a modular and gate," *Molecular Systems Biology*, vol. 3, p. 133, 2007.

[39] A. Tamsir, J. J. Tabor, and C. A. Voigt, "Robust multicellular computing using genetically encoded NOR gates and chemical "wires"," *Nature*, vol. 469, no. 7329, pp. 212–215, 2010.

[40] L. Pasotti, M. Quattrocelli, D. Galli, M. G. Cusella de Angelis, and P. Magni, "Multiplexing and demultiplexing logic functions for computing signal processing tasks in synthetic biology," *Biotechnology Journal*, vol. 6, no. 7, pp. 784–795, 2011.

[41] B. Wang and M. Buck, "Customizing cell signaling using engineered genetic logic circuits," *Trends in Microbiology*, vol. 20, no. 8, pp. 376–384, 2012.

[42] T. S. Gardner, C. R. Cantor, and J. J. Collins, "Construction of a genetic toggle switch in *Escherichia coli*," *Nature*, vol. 403, no. 6767, pp. 339–342, 2000.

[43] M. B. Elowitz and S. Leibier, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.

[44] J. Stricker, S. Cookson, M. R. Bennett, W. H. Mather, L. S. Tsimring, and J. Hasty, "A fast, robust and tunable synthetic gene oscillator," *Nature*, vol. 456, no. 7221, pp. 516–519, 2008.

[45] T. Danino, O. Mondragón-Palomino, L. Tsimring, and J. Hasty, "A synchronized quorum of genetic clocks," *Nature*, vol. 463, no. 7279, pp. 326–330, 2010.

[46] S. Basu, R. Mehreja, S. Thiberge, M. T. Chen, and R. Weiss, "Spatiotemporal control of gene expression with pulse-generating networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 17, pp. 6355–6360, 2004.

[47] S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss, "A synthetic multicellular system for programmed pattern formation," *Nature*, vol. 434, no. 7037, pp. 1130–1134, 2005.

[48] J. J. Tabor, H. Salis, Z. B. Simpson et al., "A synthetic genetic edge detection program," *Cell*, vol. 137, no. 7, pp. 1272–1281, 2009.

[49] A. E. Friedland, T. K. Lu, X. Wang, D. Shi, G. Church, and J. J. Collins, "Synthetic gene networks that count," *Science*, vol. 324, no. 5931, pp. 1199–1202, 2009.

[50] R. Daniel, J. R. Rubens, R. Sarpeshkar, and T. K. Lu, "Synthetic analog computation in living cells," *Nature*, vol. 497, no. 7451, pp. 619–623, 2013.

[51] C. J. Paddon, P. J. Westfall, D. J. Pitera et al., "High-level semi-synthetic production of the potent antimalarial artemisinin," *Nature*, vol. 496, no. 7446, pp. 528–532, 2013.

[52] K. de Mora, N. Joshi, B. L. Balint, F. B. Ward, A. Elfick, and C. E. French, "A pH-based biosensor for detection of arsenic in drinking water," *Analytical and Bioanalytical Chemistry*, vol. 400, no. 4, pp. 1031–1039, 2011.

[53] C. E. French, K. de Mora, N. Joshi, A. Elfick, J. Haseloff, and J. Ajioka, "Synthetic biology and the art of biosensor design," in *Institute of Medicine (US) Forum on Microbial Threats. The Science and Applications of Synthetic and Systems Biology: Workshop Summary*, National Academies Press, Washington, DC, USA, 2011.

[54] A. J. Wargacki, E. Leonard, M. N. Win et al., "An engineered microbial platform for direct biofuel production from brown macroalgae," *Science*, vol. 335, no. 6066, pp. 308–313, 2012.

[55] F. Zhang, J. M. Carothers, and J. D. Keasling, "Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids," *Nature Biotechnology*, vol. 30, no. 4, pp. 354–359, 2012.

[56] L. Serrano, "Synthetic biology: promises and challenges," *Molecular Systems Biology*, vol. 3, article 158, 2007.

[57] A. P. Arkin, "A wise consistency: engineering biology for conformity, reliability, predictability," *Current Opinion in Chemical Biology*, vol. 17, no. 6, pp. 893–901, 2013.

[58] T. K. Lu, A. S. Khalil, and J. J. Collins, "Next-generation synthetic gene networks," *Nature Biotechnology*, vol. 27, no. 12, pp. 1139–1150, 2009.

[59] Y. Yokobayashi, R. Weiss, and F. H. Arnold, "Directed evolution of a genetic circuit," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16587–16591, 2002.

[60] S. Kosuri, D. B. Goodman, G. Cambray et al., "Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 34, pp. 14024–14029, 2013.

[61] F. Chizzolini, M. Forlin, D. Cecchi, and S. S. Mansy, "Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength," *ACS Synthetic Biology*, 2013.

[62] F. Ceroni, S. Furini, E. Giordano, and S. Cavalcanti, "Rational design of modular circuits for gene transcription: a test of the bottom-up approach," *Journal of Biological Engineering*, vol. 4, article 14, 2010.

[63] D. Na, S. M. Yoo, H. Chung, H. Park, J. H. Park, and S. Y. Lee, "Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs," *Nature Biotechnology*, vol. 31, no. 2, pp. 170–174, 2013.

[64] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[65] M. H. Medema, R. Van Raaphorst, E. Takano, and R. Breitling, "Computational tools for the synthetic design of biochemical pathways," *Nature Reviews Microbiology*, vol. 10, no. 3, pp. 191–202, 2012.

[66] J. Ang, E. Harris, B. J. Hussey, R. Kil, and D. R. McMillen, "Tuning response curves for synthetic biology," *ACS Synthetic Biology*, vol. 2, no. 10, pp. 547–567, 2013.

[67] N. Crook and H. S. Alper, "Model-based design of synthetic, biological systems," *Chemical Engineering Science*, vol. 103, pp. 2–11, 2013.

[68] Y. Cai, M. L. Wilson, and J. Peccoud, "GenoCAD for iGEM: a grammatical approach to the design of standard-compliant constructs," *Nucleic Acids Research*, vol. 38, no. 8, pp. 2637–2644, 2010.

[69] W. Kammerer, U. Deuschle, R. Gentz, and H. Bujard, "Functional dissection of *Escherichia coli* promoters: information in the transcribed region is involved in late steps of the overall process," *The EMBO Journal*, vol. 5, no. 11, pp. 2995–3000, 1986.

[70] S. Leirmo and R. L. Gourse, "Factor-independent activation of *Escherichia coli* rRNA transcription. I. Kinetic analysis of the roles of the upstream activator region and supercoiling on transcription of the rrnB P1 promoter in vitro," *Journal of Molecular Biology*, vol. 220, no. 3, pp. 555–568, 1991.

[71] T. Caramori and A. Galizzi, "The UP element of the promoter for the flagellin gene, hag, stimulates transcription from both SigD- and SigA-dependent promoters in Bacillus subtilis," *Molecular and General Genetics*, vol. 258, no. 4, pp. 385–388, 1998.

[72] S. T. Estrem, T. Gaal, W. Ross, and R. L. Gourse, "Identification of an UP element consensus sequence for bacterial promoters," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 17, pp. 9761–9766, 1998.

[73] J. H. Davis, A. J. Rubin, and R. T. Sauer, "Design, construction and characterization of a set of insulated bacterial promoters," *Nucleic Acids Research*, vol. 39, no. 3, pp. 1131–1141, 2011.

[74] H. Alper, C. Fischer, E. Nevoigt, and G. Stephanopoulos, "Tuning genetic control through promoter engineering," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12678–12683, 2005.

[75] S. T. Estrem, W. Ross, T. Gaal et al., "Bacterial promoter architecture: Subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase α subunit," *Genes and Development*, vol. 13, no. 16, pp. 2134–2147, 1999.

[76] W. Ross, A. Ernst, and R. L. Gourse, "Fine structure of *E. coli* RNA polymerase-promoter interactions: α subunit binding to the UP element minor groove," *Genes and Development*, vol. 15, no. 5, pp. 491–506, 2001.

[77] C. L. Chan and C. A. Gross, "The anti-initial transcribed sequence, a portable sequence that impedes promoter escape, requires σ70 for function," *The Journal of Biological Chemistry*, vol. 276, no. 41, pp. 38201–38209, 2001.

[78] L. Martin, A. Che, and D. Endy, "Gemini, a bifunctional enzymatic and fluorescent reporter of gene expression," *PLoS ONE*, vol. 4, no. 11, Article ID e7569, 2009.

[79] M. Hajimorad, P. R. Gray, and J. D. Keasling, "A framework and model system to investigate linear system behavior in *Escherichia coli*," *Journal of Biological Engineering*, vol. 5, article 3, 2011.

[80] V. A. Rhodius and V. K. Mutalik, "Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, sigmaE," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 2854–2859, 2010.

[81] V. K. Mutalik, J. C. Guimaraes, G. Cambray et al., "Precise and reliable gene expression via standard transcription and translation initiation elements," *Nature Methods*, vol. 10, no. 4, pp. 354–360, 2013.

[82] D. Na, S. Lee, and D. Lee, "Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes," *BMC Systems Biology*, vol. 4, article 71, 2010.

[83] B. Reeve, T. Hargest, C. Gilbert, and T. Ellis, "Predicting translation initiation rates for designing synthetic biology," *Frontiers in Bioengineering and Biotechnology*, vol. 2, article 1, 2014.

[84] A. Espah Borujeni, A. S. Channarasappa, and H. M. Salis, "Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites," *Nucleic Acids Research*, vol. 42, no. 4, pp. 2646–2659, 2014.

[85] G. Pothoulakis, F. Ceroni, B. Reeve, and T. Ellis, "The Spinach RNA aptamer as a characterization tool for synthetic biology," *ACS Synthetic Biology*, vol. 3, 182, no. 3, p. 187, 2014.

[86] C. Bi, P. Su, J. Müller et al., "Development of a broad-host synthetic biology toolbox for ralstonia eutropha and its application to engineering hydrocarbon biofuel production," *Microbial Cell Factories*, vol. 12, article 107, 2013.

[87] F. F. Nowroozi, E. E. K. Baidoo, S. Ermakov et al., "Metabolic pathway optimization using ribosome binding site variants and combinatorial gene assembly," *Applied Microbiology and Biotechnology*, vol. 98, no. 4, pp. 1567–1581, 2014.

[88] M. Welch, S. Govindarajan, J. E. Ness et al., "Design parameters to control synthetic gene expression in *Escherichia coli*," *PLoS ONE*, vol. 4, no. 9, Article ID e7002, 2009.

[89] C. Gustafsson, J. Minshull, S. Govindarajan, J. Ness, A. Villalobos, and M. Welch, "Engineering genes for predictable protein expression," *Protein Expression and Purification*, vol. 83, no. 1, pp. 37–46, 2012.

[90] B. K. S. Chung and D. Y. Lee, "Computational codon optimization of synthetic gene for protein expression," *BMC Systems Biology*, vol. 6, p. 134, 2012.

[91] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, "Coding-sequence determinants of expression in *Escherichia coli*," *Science*, vol. 324, no. 5924, pp. 255–258, 2009.

[92] H. G. Menzella, "Comparison of two codon optimization strategies to enhance recombinant protein production in *Escherichia coli*," *Microbial Cell Factories*, vol. 10, article 15, 2011.

[93] C. Gustafsson, S. Govindarajan, and J. Minshull, "Codon bias and heterologous protein expression," *Trends in Biotechnology*, vol. 22, no. 7, pp. 346–353, 2004.

[94] M. Graf, T. Schoedl, and R. Wagner, "Rationales of gene design and de novo gene construction," in *Systems Biology and*

Synthetic Biology, P. Fu and S. Panke, Eds., pp. 411–438, John Wiley& Sons, Hoboken, NJ, USA, 2009.

[95] P. M. Sharp and W. H. Li, "The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[96] D. B. Goodman, G. M. Church, and S. Kosuri, "Causes and effects of N-terminal codon bias in bacterial genes," *Science*, vol. 342, no. 6157, pp. 475–479, 2013.

[97] C. Elena, P. Ravasi, M. E. Castelli, S. Peiru, and H. G. Menzella, "Expression of codon optimized genes in microbial systems: current industrial applications and perspectives," *Frontiers in Microbiology*, vol. 5, article 21, 2014.

[98] G. L. Rosano and E. A. Ceccarelli, "Recombinant protein expression in *Escherichia coli*: advances and challenges," *Frontiers in Microbiology*, vol. 5, p. 172, 2014.

[99] S. C. Sleight, B. A. Bartley, J. A. Lieviant, and H. M. Sauro, "Designing and engineering evolutionary robust genetic circuits," *Journal of Biological Engineering*, vol. 4, p. 12, 2010.

[100] L. Pasotti, S. Zucca, M. Lupotto, M. G. Cusella De Angelis, and P. Magni, "Characterization of a synthetic bacterial self-destruction device for programmed cell death and for recombinant proteins release," *Journal of Biological Engineering*, vol. 5, article 8, 2011.

[101] J. R. Kelly, *Tools and reference standards supporting the engineering and evolution of synthetic biological systems [Ph.D. thesis]*, Massachusetts Institute of Technology, 2008.

[102] A. Levin-Karp, U. Barenholz, T. Bareia et al., "Quantifying translational coupling in *E. coli* synthetic operons using RBS modulation and fluorescent reporters," *ACS Synthetic Biology*, vol. 2, no. 6, pp. 327–336, 2013.

[103] S. Zucca, L. Pasotti, G. Mazzini, M. G. Cusella De Angelis, and P. Magni, "Characterization of an inducible promoter in different DNA copy number conditions," *BMC Bioinformatics*, vol. 13, no. 4, article S11, 2012.

[104] N. J. Guido, X. Wang, D. Adalsteinsson et al., "A bottom-up approach to gene regulation," *Nature*, vol. 439, no. 7078, pp. 856–860, 2006.

[105] Y. Dublanche, K. Michalodimitrakis, N. Kümmerer, M. Foglierini, and L. Serrano, "Noise in transcription negative feedback loops: simulation and experimental analysis," *Molecular Systems Biology*, vol. 2, article 41, 2006.

[106] T. S. Lee, R. A. Krupa, F. Zhang et al., "BglBrick vectors and datasheets: a synthetic biology platform for gene expression," *Journal of Biological Engineering*, vol. 5, article 12, 2011.

[107] S. Zucca, L. Pasotti, N. Politi, M. G. Cusella De Angelis, and P. Magni, "A standard vector for the chromosomal integration and characterization of BioBrick parts in *Escherichia coli*," *Journal of Biological Engineering*, vol. 7, no. 1, article 12, 2013.

[108] C. Solem and P. R. Jensen, "Modulation of gene expression made easy," *Applied and Environmental Microbiology*, vol. 68, no. 5, pp. 2397–2403, 2002.

[109] L. O. Ingram and T. Conway, "Expression of different levels of ethanologenic enzymes from *Zymomonas mobilis* in recombinant strains of *Escherichia coli*," *Applied and Environmental Microbiology*, vol. 54, no. 2, pp. 397–404, 1988.

[110] A. Martinez, S. W. York, L. P. Yomano et al., "Biosynthetic burden and plasmid burden limit expression of chromosomally integrated heterologous genes (pdc, adhB) in *Escherichia coli*," *Biotechnology Progress*, vol. 15, no. 5, pp. 891–897, 1999.

[111] T. Ellis, X. Wang, and J. J. Collins, "Diversity-based, model-guided construction of synthetic gene networks with predicted functions," *Nature Biotechnology*, vol. 27, no. 5, pp. 465–471, 2009.

[112] K. Temme, D. Zhao, and C. A. Voigt, "Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 18, pp. 7085–7090, 2012.

[113] N. Politi, L. Pasotti, S. Zucca et al., "Half-life measurements of chemical inducers for recombinant gene expression," *Journal of Biological Engineering*, vol. 8, article 5, 2014.

[114] R. Weiss, *Cellular computation and communications using engineered genetic regulatory networks [Ph.D. thesis]*, Massachusetts Institute of Technology, 2001.

[115] H. H. Wang, F. J. Isaacs, P. A. Carr et al., "Programming cells by multiplex genome engineering and accelerated evolution," *Nature*, vol. 460, no. 7257, pp. 894–898, 2009.

[116] B. F. Pfleger, D. J. Pitera, C. D. Smolke, and J. D. Keasling, "Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes," *Nature Biotechnology*, vol. 24, no. 8, pp. 1027–1032, 2006.

[117] N. Sawada, T. Sakaki, S. Kitanaka, K. Takeyama, S. Kato, and K. Inouye, "Enzymatic properties of human 25-hydroxyvitamin D3 1α-hydroxylase. Coexpression with adrenodoxin and NADPH-adrenodoxin reductase in *Escherichia coli*," *European Journal of Biochemistry*, vol. 265, no. 3, pp. 950–956, 1999.

[118] J. T. Kittleson, S. Cheung, and J. C. Anderson, "Rapid optimization of gene dosage in *E. coli* using DIAL strains," *Journal of Biological Engineering*, vol. 5, article 10, 2011.

[119] C. N. Santos, D. D. Regitsky, and Y. Yoshikuni, "Implementation of stable and complex biological systems through recombinase-assisted genome engineering," *Nature Communications*, vol. 4, article 2503, 2013.

[120] L. P. Yomano, S. W. York, S. Zhou, K. T. Shanmugam, and L. O. Ingram, "Re-engineering *Escherichia coli* for ethanol production," *Biotechnology Letters*, vol. 30, no. 12, pp. 2097–2103, 2008.

[121] K. Ohta, D. S. Beall, J. P. Mejia, K. T. Shanmugam, and L. O. Ingram, "Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II," *Applied and Environmental Microbiology*, vol. 57, no. 4, pp. 893–900, 1991.

[122] P. C. Turner, L. P. Yomano, L. R. Jarboe et al., "Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis* pdc and *adhB* genes," *Journal of Industrial Microbiology & Biotechnology*, vol. 39, no. 4, pp. 629–639, 2012.

[123] K. E. J. Tyo, P. K. Ajikumar, and G. Stephanopoulos, "Stabilized gene duplication enables long-term selection-free heterologous pathway expression," *Nature Biotechnology*, vol. 27, no. 8, pp. 760–765, 2009.

[124] L. Qi, R. E. Haurwitz, W. Shao, J. A. Doudna, and A. P. Arkin, "RNA processing enables predictable programming of gene expression," *Nature Biotechnology*, vol. 30, no. 10, pp. 1002–1006, 2012.

[125] C. Lou, B. Stanton, Y. Chen, B. Munsky, and C. A. Voigt, "Ribozyme-based insulator parts buffer synthetic circuits from genetic context," *Nature Biotechnology*, vol. 30, no. 11, pp. 1137–1142, 2012.

[126] D. Del Vecchio, "A control theoretic framework for modular analysis and design of biomolecular networks," *Annual Reviews in Control*, vol. 37, pp. 333–345, 2013.

[127] B. Li and L. You, "Predictive power of cell-to-cell variability," *Quantitative Biology*, vol. 1, no. 2, pp. 131–139, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 761907, 9 pages
http://dx.doi.org/10.1155/2014/761907

Research Article

Effects of Maximal Sodium and Potassium Conductance on the Stability of Hodgkin-Huxley Model

Yue Zhang,[1] Kuanquan Wang,[1] Yongfeng Yuan,[1] Dong Sui,[1] and Henggui Zhang[1,2]

[1] *Biocomputing Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[2] *School of Physics & Astronomy, University of Manchester, Manchester, UK*

Correspondence should be addressed to Kuanquan Wang; wangkq@hit.edu.cn

Received 13 February 2014; Revised 12 June 2014; Accepted 15 June 2014; Published 3 July 2014

Academic Editor: Rui Jiang

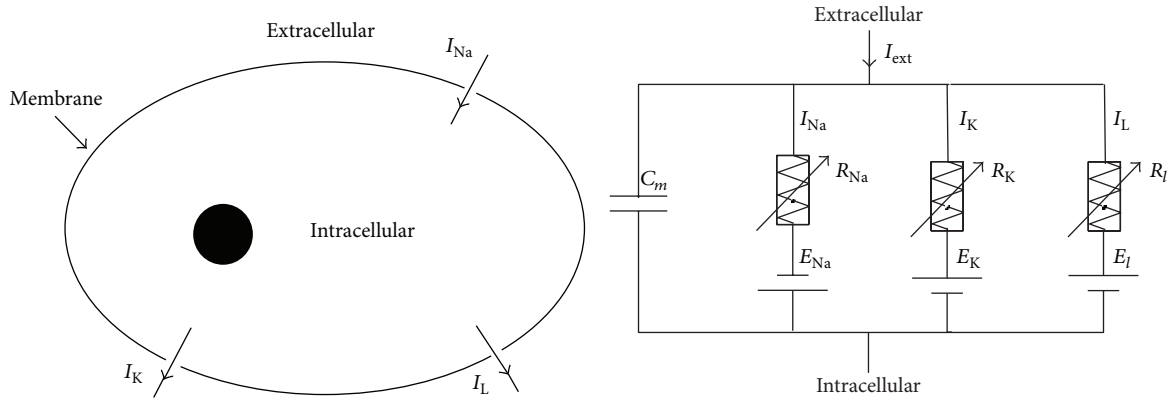
Hodgkin-Huxley (HH) equation is the first cell computing model in the world and pioneered the use of model to study electrophysiological problems. The model consists of four differential equations which are based on the experimental data of ion channels. Maximal conductance is an important characteristic of different channels. In this study, mathematical method is used to investigate the importance of maximal sodium conductance $\overline{g}_{Na}$ and maximal potassium conductance $\overline{g}_{K}$. Applying stability theory, and taking $\overline{g}_{Na}$ and $\overline{g}_{K}$ as variables, we analyze the stability and bifurcations of the model. Bifurcations are found when the variables change, and bifurcation points and boundary are also calculated. There is only one bifurcation point when $\overline{g}_{Na}$ is the variable, while there are two points when $\overline{g}_{K}$ is variable. The ($\overline{g}_{Na}$, $\overline{g}_{K}$) plane is partitioned into two regions and the upper bifurcation boundary is similar to a line when both $\overline{g}_{Na}$ and $\overline{g}_{K}$ are variables. Numerical simulations illustrate the validity of the analysis. The results obtained could be helpful in studying relevant diseases caused by maximal conductance anomaly.

1. Introduction

Hodgkin-Huxley (HH) equation is created on the foundation of huge experimental data of sodium and potassium channels by Hodgkin and Huxley who are both excellent biology scientists and had long engaged in nerve conduction research. In about 1952, they took squid giant axon as experiment subject and continuously published four papers describing the electrical excitation of this kind of cell [1–4]. In their experiment, all the ion channels were divided into three types, sodium channel, potassium channel, and the others. Now we know there are many ion channels on the cell membrane, such as I_{Na}, I_{Kr}, I_{Ks}, I_{NaCa}, I_{K1}, I_{CaL}, I_{Ca}, I_{to}, I_{NaK}, I_{NaL}, and I_{KATP} [5–7]. However the discovery of sodium and potassium channels was marvelous at that time. Experimental data was obtained by voltage-clamp technology, while the patch-clamp technology is widely used at present. On this basis, a four-dimensional ordinary differential equation set, called HH model, was proposed, which was autonomous and contained intricate transcendental equations.

The work of Hodgkin and Huxley was recognized as excellent achievement and with significant contribution to the development of electrophysiology. It is the basis of the subsequent models of ion channels. Not only was the HH model consistent with the obtained experiment data accurately, but also it could precisely simulate the change of action potential. The model discovered the relationship of transmembrane potential and current and maximum conductance of ions. This made it possible to research the character of ion channel with mathematical methods. In 1960, Professor Nobel who pioneered the cardiac electrophysiology simulations applied HH model to myocardial cell and got the famous Purkinje fiber cell model [8], which was the first computing myocardial cell model. From then on, HH model was broadly applied to almost all kinds of cardiac cells such as atrial muscle cell model [9] and sinoatrial node cell model [10]. HH model laid the cornerstone of computing electrophysiology. Even today, a large part of electrophysiological models are created on the foundation of HH model. Verkerk's sinoatrial model [11], Butters's atrial model [12], O'Hara's ventricular model [13], and Li's Purkinje cell model [14], and so forth, all belong to HH model type.

Because of the importance of HH model, the stability has long attracted the researcher's attention. Hassard et al.



(a) Electrophysiological activities of cell

(b) Electrical equivalent circuit of HH model

Figure 1: The electrophysiological process and equivalent circuit of neuron.

were the earlier researchers caring about the bifurcation phenomenon of HH model. And they indicated that bifurcation would occur at the equilibrium points when the external current i_{ext} changed which was injected into the neuron from microelectrode [15]. Stable and unstable solutions of the model with regard to i_{ext} were analyzed by Rinzel and Miller, and the influence of temperature was also discussed [16]. Two stable steady states were found by Aihara and Matsumoto [17]; when the two states existed, the bifurcation structure was complex, which included a stable limit cycle, two unstable equilibrium points, and one asymptotically stable equilibrium point. Guckenheimer and Labouriau investigated the influence of i_{ext} and potassium ion potential V_K on the bifurcations of the model [18]. Bedrov et al. gave the relationship between the numbers of negative slope regions and presented some results about the possible bifurcation giving rise to maximal sodium conductance $\overline{g}_{Na}$ and maximal potassium conductance $\overline{g}_K$ [19, 20]. Fukai and his fellows examined the structure of the model's bifurcations produced by i_{ext} and one of the other parameters [21]. Taking leakage conductance g_l and sodium channel effective bias voltage $\overline{V}_m$ as parameters, Terada et al. analyzed bifurcation in Hodgkin-Huxley model for muscles of frogs [22].

Wang et al. [23–26] did a lot of research on the stability of HH model. They studied the bifurcations caused by leakage conductance $\overline{g}_l$ and sodium ions antielectromotive force when ELF external electric field was considered. The stability and bifurcation control were analyzed and controllers were designed. Bifurcation in HH model exposed to DC electric fields was investigated in detail.

Bifurcation means qualitative changes in the solution structure of a dynamic system when the parameters vary. From analyzing the bifurcation, we can get the effects of the parameters. Further, changing the corresponding parameters, we could make the solution into an ideal condition. Bifurcation is an important branch in mathematics and applied to much different field [27–29]. In recent years, it is also widely studied in electrophysiology. Indeed, there are many diseases having close relations with bifurcations, such as Parkinson's, epilepsy, and pathological heart rhythms [30].

In the past, for HH model, external current i_{ext} and leakage conductance $\overline{g}_l$ have been most investigated, because they were easily measured. The sodium current is the contributor which leads to depolarization of the neuron while it is potassium current that plays the major role of repolarization. However, $\overline{g}_{Na}$ and $\overline{g}_K$ are seldom taken into consideration to analyze the stability of model, as the relevant data is not abundant. In this study, the effects of $\overline{g}_{Na}$ and $\overline{g}_K$ on the stability and bifurcations of the model will be discussed, respectively, and collectively. And we will give the critical points of $\overline{g}_{Na}$ and $\overline{g}_K$ when they play the role separately, and the critical boundaries in $\overline{g}_{Na}$-$\overline{g}_K$ plane will be provided when together. Simulation results demonstrate the validity of the theoretic analysis.

The rest of the paper is organized as follows. The HH equations are introduced in detail in Section 2. In Section 3 we analyze the effects of $\overline{g}_{Na}$ and $\overline{g}_K$ on the model and calculate the bifurcation points (line). Finally, discussion and conclusion are presented in Section 4.

2. Hodgkin-Huxley Equations

HH model was proposed on the foundation of ion channels. The electrophysiological activities of a cell are shown in Figure 1(a). The gray circle is membrane, which ensures orderly biochemical reaction. I_{Na}, I_K, and I_L are the ion currents corresponding to respective channels on the membrane. When an electrical stimulation makes the sodium channels open, a large number of Na^+ flow inward, forming current I_{Na}, resulting in the rise of transmembrane potential. The open of potassium channels makes a large outflux of K^+, creating the current I_K and the reduction of potential. The model is comprised of four autonomous ordinary differential equations to describe the electrophysiological activities of cell shown in Figure 1(a). In the model, membrane is taken as a constant capacitance and the ion channels are seen as variable resistances. Figure 1(b) shows the equivalent circuit in detail, in which $R_{Na} = 1/g_{Na}$, $R_K = 1/g_K$, and $R_l = 1/g_l$. R_{Na} and R_K vary with time.

The equations were obtained according to electrical formulas and experimental data, which are shown as follows:

$$\begin{aligned}\frac{dV}{dt} &= \frac{1}{C_M}\left[i_{\text{ext}} - \overline{g}_{\text{Na}} m^3 h\left(V - V_{\text{Na}}\right)\right.\\ &\qquad \left. -\overline{g}_{\text{K}} n^4\left(V - V_{\text{K}}\right) - \overline{g}_l\left(V - V_l\right)\right],\\ \frac{dm}{dt} &= \alpha_m(V)(1-m) - \beta_m(V)\, m,\\ \frac{dh}{dt} &= \alpha_h(V)(1-h) - \beta_h(V)\, h,\\ \frac{dn}{dt} &= \alpha_n(V)(1-n) - \beta_n(V)\, n,\end{aligned} \tag{1}$$

where

$$\begin{aligned}\alpha_m(V) &= \frac{0.1\,(V - 25.0)}{1 - \exp\left[-(V - 25.0)/10\right]},\\ \beta_m(V) &= 4.0\exp\left(-\frac{V}{18.0}\right),\\ \alpha_h(V) &= 0.07\exp\left(-\frac{V}{20.0}\right),\\ \beta_h(V) &= \frac{1}{1 + \exp\left[-(V - 30.0)/10\right]},\\ \alpha_n(V) &= \frac{0.01\,(V - 10.0)}{1 - \exp\left[-(V - 10.0)/10\right]},\\ \beta_n(V) &= 0.125\exp\left(-\frac{V}{80.0}\right).\end{aligned} \tag{2}$$

In these equations, V is the transmembrane potential. $0 \le m \le 1$ and $0 \le h \le 1$ are the gating variables indicating activation and inactivation of sodium ion current, respectively. $0 \le n \le 1$ is the gating variable showing activation of potassium ion current. $\overline{g}_{\text{Na}}$, $\overline{g}_{\text{K}}$, and $\overline{g}_l$ represent the maximal conductance of corresponding currents. $C_m = 1.0\,\mu\text{F/cm}^2$ is membrane capacitance. i_{ext} is the current injected into the neuron. In our paper, we suppose $i_{\text{ext}} = 0$ and $\overline{g}_{\text{Na}} = 120\,\text{mS/cm}^2$, $\overline{g}_{\text{K}} = 36\,\text{mS/cm}^2$, and $\overline{g}_l = 0.3\,\text{mS/cm}^2$, which are the ideal experimental data.

3. Stability Analysis of HH Model

Stability is one of a model's important properties. If the model is stable, it will reach a rest state at last. Otherwise, periodic phenomenon or chaos may appear. To analyze an ordinary differential system, equilibrium points are one of its most important aspects, which may be the final state of the system. Suppose (V_*, m_*, h_*, n_*) is the equilibrium points of the model. So it should make the right side of (1) equal to zero. That is,

$$\begin{aligned}&i_{\text{ext}} - \overline{g}_{\text{Na}} m_*^3 h_*\left(V_* - V_{\text{Na}}\right) - \overline{g}_{\text{K}} n_*^4\left(V_* - V_{\text{K}}\right)\\ &\quad - \overline{g}_l\left(V_* - V_l\right) = 0,\\ &\alpha_m(V_*)(1 - m_*) - \beta_m(V_*)\, m_* = 0,\\ &\alpha_h(V_*)(1 - h_*) - \beta_h(V_*)\, h_* = 0,\\ &\alpha_n(V_*)(1 - n_*) - \beta_n(V_*)\, n_* = 0.\end{aligned} \tag{3}$$

Then the linearization of (1) around the equilibrium could be obtained as follows:

$$\begin{aligned}\frac{dV}{dt} &= J_{11}V + J_{12}m + J_{13}h + J_{14}n,\\ \frac{dm}{dt} &= J_{21}V + J_{22}m,\\ \frac{dh}{dt} &= J_{31}V + J_{33}h,\\ \frac{dn}{dt} &= J_{41}V + J_{44}n,\end{aligned} \tag{4}$$

where

$$\begin{aligned}J_{11} &= -\frac{\overline{g}_{\text{Na}} m_*^3 h_* + \overline{g}_{\text{K}} n_*^4 + \overline{g}_l}{C_M},\\ J_{12} &= -\frac{3\overline{g}_{\text{Na}} m_*^2 h_*\left(V_* - V_{\text{Na}}\right)}{C_M},\\ J_{13} &= -\frac{\overline{g}_{\text{Na}} m_*^3\left(V_* - V_{\text{Na}}\right)}{C_M},\\ J_{14} &= -\frac{4\overline{g}_{\text{K}} n_*^3\left(V_* - V_{\text{K}}\right)}{C_M},\\ J_{21} &= \frac{2m_*}{9\exp\left(V_*/18.0\right)}\\ &\quad - \left\{\frac{0.1\,(V_* - 25.0)\exp\left[-(V_* - 25.0)/10\right]}{\left[1 - \exp\left(-(V_* - 25)/10\right)\right]^2}\right.\\ &\quad \left. - \frac{0.1}{1 - \exp\left(-(V_* - 25)/10\right)}\right\}(1 - m_*),\\ J_{22} &= \frac{0.1\,(V_* - 25.0)}{1 - \exp\left[-(V_* - 25.0)/10\right]} - 4.0\exp\left(-\frac{V_*}{18.0}\right),\\ J_{31} &= -\frac{7\exp\left(V_*/20.0\right)}{2000}(1 - h_*)\\ &\quad - \frac{\exp\left(-(V_* - 30.0)/10\right)h_*}{10\left[1 + \exp\left(-(V_* - 30.0)/10\right)\right]^2},\\ J_{33} &= 0.07\exp\left(-\frac{V_*}{20.0}\right) - \frac{1}{1 + \exp\left(-(V_* - 30.0)/10\right)},\\ J_{41} &= \frac{n_*\exp\left(V_*/80.0\right)}{640}\\ &\quad + \left\{\frac{1}{100\left[1 - \exp\left(-(V_* - 10.0)/10\right)\right]}\right.\\ &\quad \left. - \frac{0.01\exp\left[-(V_* - 10.0)/10\right](V_* - 10.0)}{10\left[1 - \exp\left(-(V_* - 10.0)/10\right)\right]^2}\right\}(1 - n_*),\\ J_{44} &= \frac{0.01\,(V_* - 10.0)}{1 - \exp\left[-(V_* - 10.0)/10\right]} - 0.125\exp\left(-\frac{V_*}{80.0}\right).\end{aligned} \tag{5}$$

We can get the eigenmatrix of (4):

$$J = \begin{pmatrix} J_{11} & J_{12} & J_{13} & J_{14} \\ J_{21} & J_{22} & 0 & 0 \\ J_{31} & 0 & J_{33} & 0 \\ J_{41} & 0 & 0 & J_{44} \end{pmatrix} \tag{6}$$

and then the characteristic equation can be obtained:

$$\lambda^4 + a\lambda^3 + b\lambda^2 + c\lambda + d = 0, \tag{7}$$

where

$$\begin{aligned} a &= -\left(J_{11} + J_{22} + J_{33} + J_{44}\right), \\ b &= J_{11}\left(J_{22} + J_{33} + J_{44}\right) + J_{22}\left(J_{33} + J_{44}\right) \\ &\quad + J_{33}J_{44} - J_{12}J_{21} - J_{13}J_{31} - J_{14}J_{41}, \\ c &= J_{12}J_{21}\left(J_{33} + J_{44}\right) + J_{13}J_{31}\left(J_{22} + J_{44}\right) \\ &\quad + J_{14}J_{41}\left(J_{22} + J_{33}\right) - J_{11}J_{22}\left(J_{33} + J_{44}\right) \\ &\quad - \left(J_{11} + J_{22}\right)J_{33}J_{44}, \\ d &= J_{11}J_{22}J_{33}J_{44} - J_{12}J_{21}J_{33}J_{44} - J_{13}J_{22}J_{31}J_{44} \\ &\quad - J_{14}J_{22}J_{33}J_{41}. \end{aligned} \tag{8}$$

According to Routh-Hurwitz criterion, if $a > 0$, $ab > c$, $d > 0$, $abc > c^2 + a^2 d$, the real parts of all the roots are minus. Otherwise, all the real parts are not negative. In the following, we will analyze the stability of the model according to this criterion.

3.1. Effect of $\overline{g}_{Na}$ on the Stability. In this section, we will investigate the influence of $\overline{g}_{\mathrm{Na}}$ on the equilibrium, stability, and bifurcations of the model. $\overline{g}_{\mathrm{Na}}$ is taken as variable, and the other parameters are all kept with desired values. Because the desired value of $\overline{g}_{\mathrm{Na}}$ is around 120 mS/cm^2, $\overline{g}_{\mathrm{Na}} \in [0, 500]$ is taken into consideration. When $\overline{g}_{\mathrm{Na}}$ changes, making the right side of (1) equal to zero, corresponding V^* can be acquired. Taking Matlab as a tool, we could obtain the relationship between $\overline{g}_{\mathrm{Na}}$ and the equilibrium point V^* shown in Figure 2.

From Figure 2, we can see that V^* changes slowly when $\overline{g}_{\mathrm{Na}} \in [0, 300]$ and increases rapidly when $\overline{g}_{\mathrm{Na}} \in [350, 500]$. This means that equilibrium points are sensitive to $\overline{g}_{\mathrm{Na}}$ when $\overline{g}_{\mathrm{Na}} \in [350, 500]$; a slight change of $\overline{g}_{\mathrm{Na}}$ may lead the model to a totally different state even though model is still stable.

Applying bifurcation theory and using the method of bisection, we can get one bifurcation point $\overline{g}^*_{\mathrm{Na}} = 212.648720656$ when $\overline{g}_{\mathrm{Na}}$ changes.

Substituting $\overline{g}^*_{\mathrm{Na}}$ into the original equation, we get the equilibrium V^*, and then substituting both $\overline{g}^*_{\mathrm{Na}}$ and V^* into eigenmatrix of (4), we can gain the eigenvalues as follows:

$$\begin{aligned} \lambda_1 &= -4.9711711484, \\ \lambda_2 &= -0.1259717048, \\ \lambda_3 &= 1.9 \times 10^{-16} - 0.3798402483i, \\ \lambda_4 &= 1.9 \times 10^{-16} + 0.3798402483i. \end{aligned} \tag{9}$$

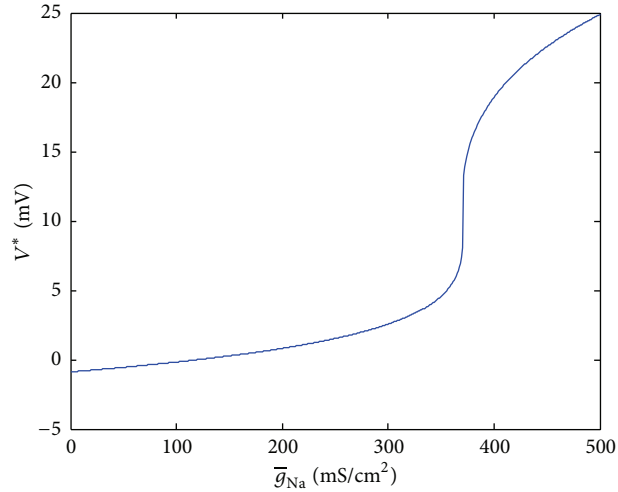


Figure 2: The relationship between $\overline{g}_{\mathrm{Na}}$ and V^*.

Here, we regard 1.9×10^{-16} as 0. With the help of computer, we can get that all the real parts of λ_i $(i = 1, 2, 3, 4)$ are negative when $\overline{g}_{\mathrm{Na}} < \overline{g}^*_{\mathrm{Na}}$. And there exist positive real parts when $\overline{g}_{\mathrm{Na}} > \overline{g}^*_{\mathrm{Na}}$. According to the stability theory, the system is stable around equilibrium when $\overline{g}_{\mathrm{Na}} \in [0, \overline{g}^*_{\mathrm{Na}})$ and it is unstable when $\overline{g}_{\mathrm{Na}} \in (\overline{g}^*_{\mathrm{Na}}, 500]$. The model undergoes Hopf bifurcation at $\overline{g}_{\mathrm{Na}} = \overline{g}^*_{\mathrm{Na}}$.

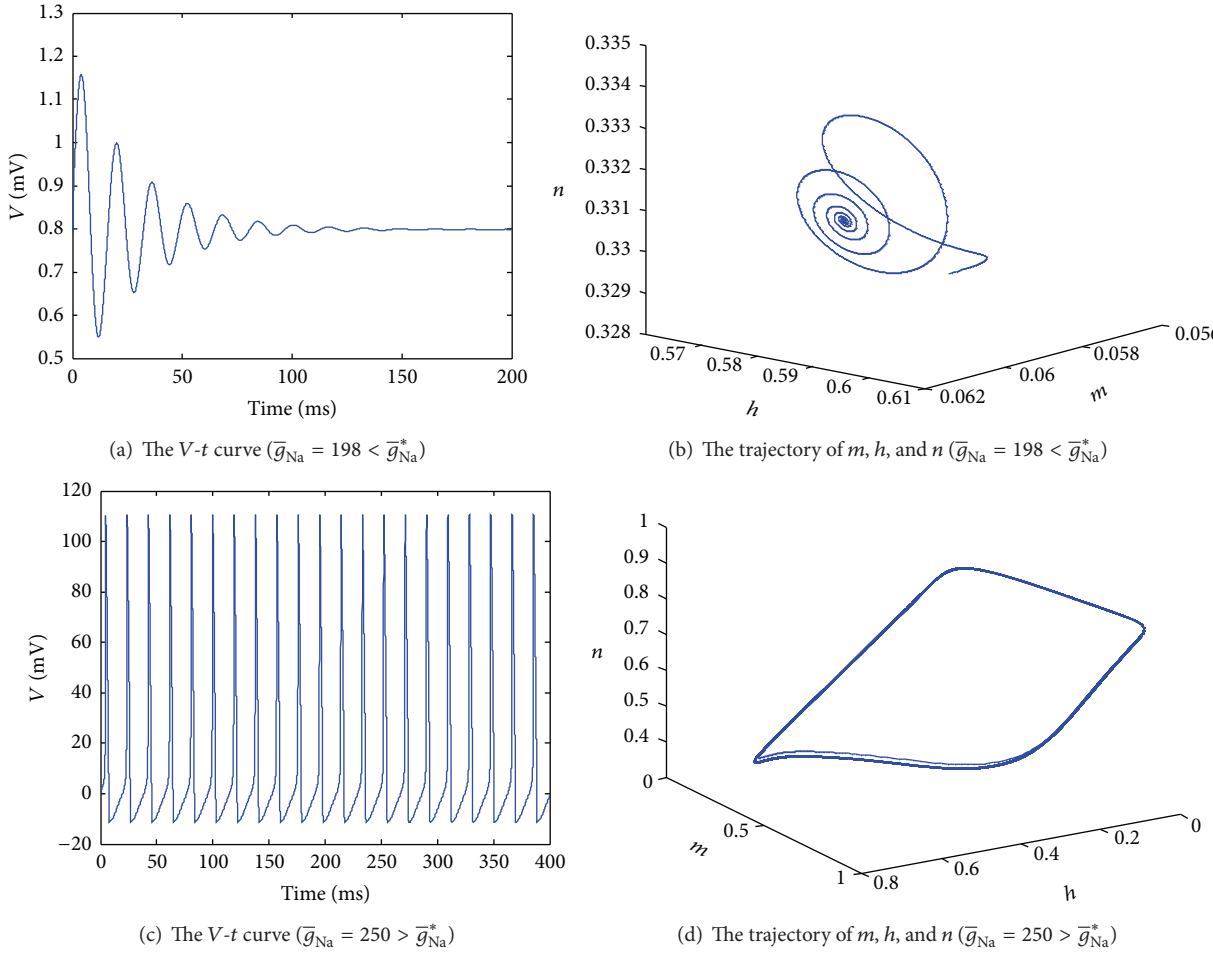
Figure 3 shows the response of V and m, h, and n to different $\overline{g}_{\mathrm{Na}}$. As the analysis above, when $\overline{g}_{\mathrm{Na}} = 198 < \overline{g}^*_{\mathrm{Na}}$, the system is stable. Figure 3(a) is the potential-time (V-t) curve, which shows that the action potential V becomes steady. Figure 3(b) displays the trajectory of gating variables m, h, and n with time. We can see that the electrophysiological activity of cell reaches an equilibrium state at last.

Figures 3(c) and 3(d) demonstrate that the system is unstable when $\overline{g}_{\mathrm{Na}} = 250 > \overline{g}^*_{\mathrm{Na}}$. Figure 3(c) is the V-t graph, from which we can see the potential changes periodically. Figure 3(d) describes the trend of m, h, and n, whose trajectory is a circle finally. Both Figures 3(c) and 3(d) imply that the system is unstable and the electrophysiological activity of cell is in a periodical state at a certain frequency.

3.2. Effect of $\overline{g}_K$ on the Model. In this part, we choose $\overline{g}_{\mathrm{K}}$ as variable and keep the other parameters with ideal values. The same method with analysis of $\overline{g}_{\mathrm{Na}}$ is taken to analyze the effect of $\overline{g}_{\mathrm{K}}$ on the equilibrium, stability, and bifurcation of HH model. $\overline{g}_{\mathrm{K}} \in [0, 200]$ is taken into consideration because the desired value of $\overline{g}_{\mathrm{K}}$ is 36.0. First, the relationship between $\overline{g}_{\mathrm{K}}$ and the equilibrium V^* is obtained in Figure 4.

From Figure 4, we can see that V^* varies rapidly when $\overline{g}_{\mathrm{K}} \in [0, 20]$ and decreases slowly when $\overline{g}_{\mathrm{K}} \in [30, 200]$. This means that equilibrium points are sensitive to $\overline{g}_{\mathrm{K}}$ when $\overline{g}_{\mathrm{K}} \in [0, 20]$. A slight change of $\overline{g}_{\mathrm{K}}$ may make the final state of model change greatly.

Using the method of bisection to calculate the eigenvalues, we can find two bifurcation points $\overline{g}^*_{\mathrm{K}1} = 3.843499029$



(a) The V-t curve ($\overline{g}_{\mathrm{Na}} = 198 < \overline{g}^*_{\mathrm{Na}}$)

(b) The trajectory of m, h, and n ($\overline{g}_{\mathrm{Na}} = 198 < \overline{g}^*_{\mathrm{Na}}$)

(c) The V-t curve ($\overline{g}_{\mathrm{Na}} = 250 > \overline{g}^*_{\mathrm{Na}}$)

(d) The trajectory of m, h, and n ($\overline{g}_{\mathrm{Na}} = 250 > \overline{g}^*_{\mathrm{Na}}$)

FIGURE 3: The response of V and m, h, and n to different $\overline{g}_{\mathrm{Na}}$.

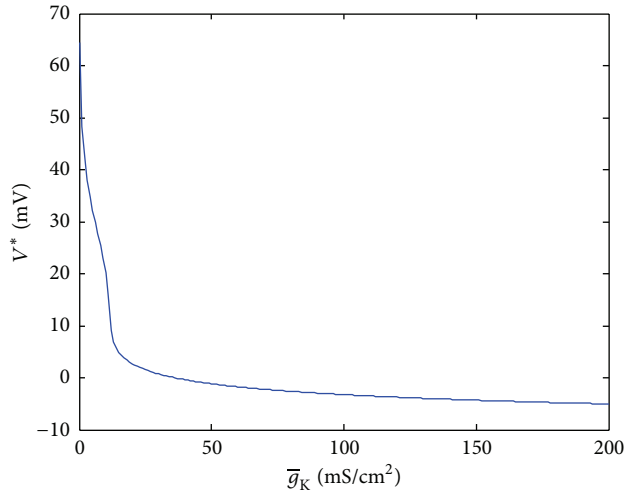
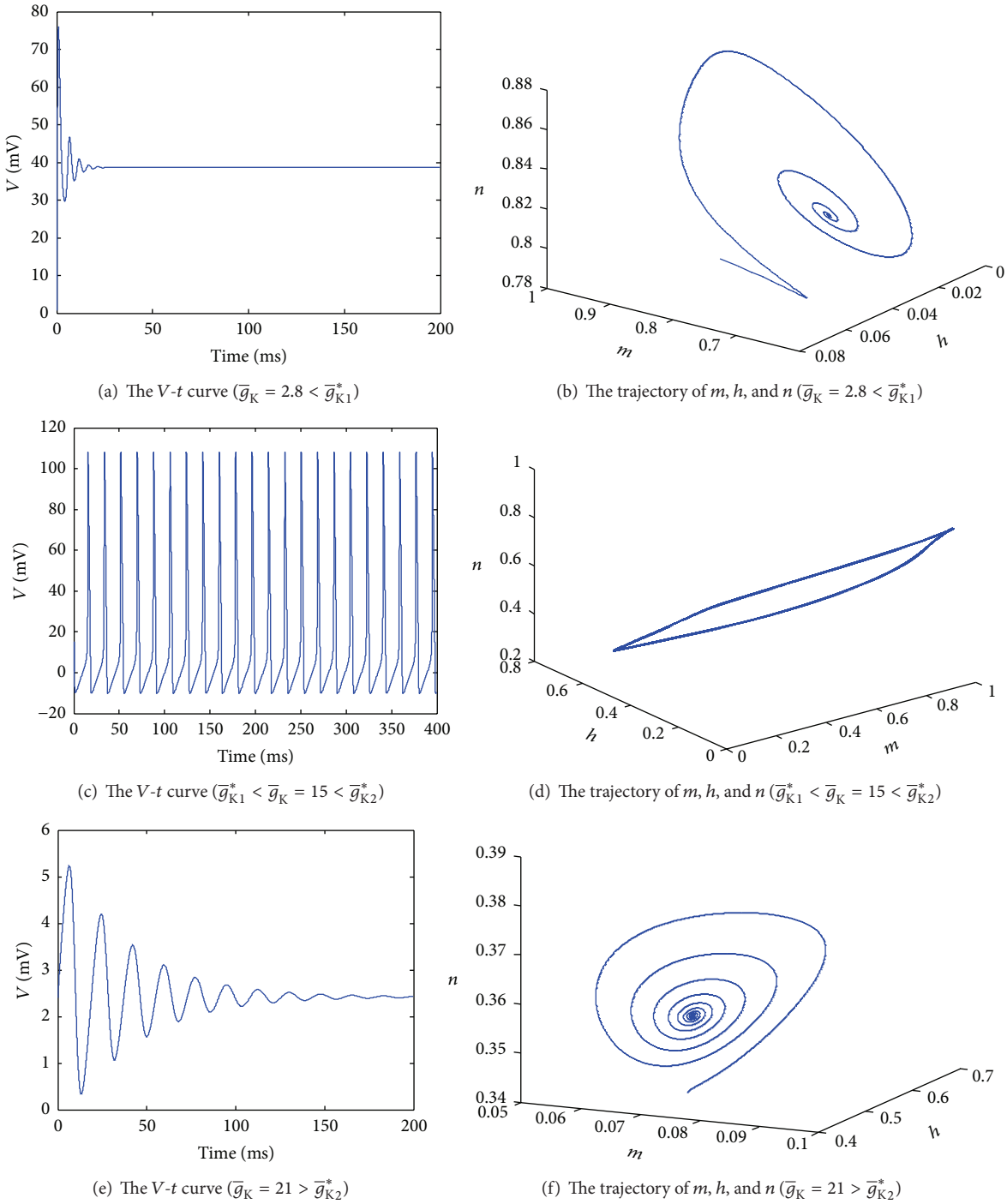


FIGURE 4: The relationship between $\overline{g}_{\mathrm{K}}$ and V^*.

and $\overline{g}^*_{\mathrm{K2}} = 19.762260771$ when $\overline{g}_{\mathrm{K}}$ varies. Substitute $\overline{g}^*_{\mathrm{K}i}$ ($i = 1, 2$) into (1), and obtain the corresponding equilibrium points V^*. Both $\overline{g}^*_{\mathrm{K}i}$ and V^* are substituted into (7), and then corresponding eigenvalues could be obtained as follows:

$$\begin{aligned}
\lambda^1_1 &= -5.3218099843,\\
\lambda^1_2 &= -0.4223840650,\\
\lambda^1_3 &= 3.2 \times 10^{-16} - 1.1305093754i,\\
\lambda^1_4 &= 3.2 \times 10^{-16} + 1.1305093754i,\\
\lambda^2_1 &= -4.5370272278,\\
\lambda^2_2 &= -0.1319002182,\\
\lambda^2_3 &= 4.2 \times 10^{-16} - 0.3436440068i,\\
\lambda^2_4 &= 4.2 \times 10^{-16} + 0.3436440068i.
\end{aligned} \tag{10}$$

Here, 3.2×10^{-16} and 4.2×10^{-16} can be approximately regarded as 0. From computing, all the real parts of eigenvalues are negative when $\overline{g}_{\mathrm{K}} \in [0, \overline{g}^*_{\mathrm{K1}}) \cup (\overline{g}^*_{\mathrm{K2}}, 200]$, and all of them are not negative when $\overline{g}_{\mathrm{K}} \in (\overline{g}^*_{\mathrm{K1}}, \overline{g}^*_{\mathrm{K2}})$. According to



(a) The V-t curve ($\overline{g}_K = 2.8 < \overline{g}^*_{K1}$)

(b) The trajectory of m, h, and n ($\overline{g}_K = 2.8 < \overline{g}^*_{K1}$)

(c) The V-t curve ($\overline{g}^*_{K1} < \overline{g}_K = 15 < \overline{g}^*_{K2}$)

(d) The trajectory of m, h, and n ($\overline{g}^*_{K1} < \overline{g}_K = 15 < \overline{g}^*_{K2}$)

(e) The V-t curve ($\overline{g}_K = 21 > \overline{g}^*_{K2}$)

(f) The trajectory of m, h, and n ($\overline{g}_K = 21 > \overline{g}^*_{K2}$)

FIGURE 5: The response of V and m, h, and n to different $\overline{g}_K$.

the stability theory, the system is stable around equilibrium when $\overline{g}_K \in [0, \overline{g}^*_{K1}) \cup (\overline{g}^*_{K2}, 200]$ and it is unstable when $\overline{g}_K \in (\overline{g}^*_{K1}, \overline{g}^*_{K2})$. The model undergoes Hopf bifurcations at $\overline{g}_K = \overline{g}^*_{Ki}$ ($i = 1, 2$). The system is from locally stable state ($\overline{g}_K \in [0, \overline{g}^*_{K1})$) to unstable state ($\overline{g}_K \in (\overline{g}^*_{K1}, \overline{g}^*_{K2})$) and becomes stable ($\overline{g}_K \in (\overline{g}^*_{K2}, 200]$) again. Responses of V and m, h, and n to different $\overline{g}_K$ are shown in Figure 5.

Figure 5 shows the response of V and m, h, and n to different $\overline{g}_K$. When $\overline{g}_K = 2.8 < \overline{g}^*_{K1}$, the system is stable. Figure 5(a) shows the trend of potential with time, from which we can see that the potential reaches a fixed value. Figure 5(b) is the trajectory of m, h, and n with time. All the gating variables also stay at fixed values (a steady point in Figure 5(b)) at last. These mean that the electrophysiological activity of cell reaches a steady state ultimately.

Figures 5(c) and 5(d) are V-t and m-h-n graphs, respectively, when $\overline{g}^*_{K1} < \overline{g}_K = 15 < \overline{g}^*_{K2}$. Figure 5(c) shows that the action potential changes in a certain period. And

Figure 5(d) describes the trajectory of m, h, and n with time, from which we can find that the shape of the trajectory is a loop. Figures 5(c) and 5(d) imply all the gating variables and potential change periodically, which means that the electrophysiological activity of cell is in a periodical state.

Figures 5(e) and 5(f) are V-t and m-h-n curves, respectively, when $\overline{g}_{\mathrm{K}} = 21 > \overline{g}^{*}_{\mathrm{K}2}$. From Figure 5(e) we can see that the potential reaches the resting state at this occasion. Figure 5(f) describes the trajectory of m, h, and n, which shows that the three variables stay at a fixed point at last. Both Figures 5(e) and 5(f) show that all the potential and gating variables no longer change with time, which implies the cell reaches the resting state finally.

3.3. Effect of $\overline{g}_{Na}$ and $\overline{g}_{K}$ on the Model. Both $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are taken as variables in this part to study the stability and bifurcation of the model when $\overline{g}_{\mathrm{Na}} \in [0, 400]$ and $\overline{g}_{\mathrm{K}} \in [0, 60]$. Keeping the other parameters with desired values, using Matlab as a tool, we get the equilibrium points first when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ both vary. Then the points are substituted into eigenmatrix of (4) and the eigenvalues of the model can be calculated. At last, Figure 6 is gained, in which all the real parts of eigenmatrix are negative if $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ belong to the pink region and positive real parts appear if $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are in white area.

From Figure 6, we can find that the upper boundary of the regions is similar to a line. With the least square method applied, the expression of the line can be gotten as $\overline{g}_{\mathrm{K}} = 0.175 \times \overline{g}_{\mathrm{Na}} - 1.675$. However, the lower boundary is not regular. According to stability theory, we can easily know that the model is stable when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are in pink region and unstable when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are in white. This means that the electrophysiological activity can reach a steady state when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are in pink region and it is periodic when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are in white. The system undergoes bifurcations when $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ are on the boundary.

4. Discussion and Conclusion

The effects of $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ on the stability and bifurcation of HH model are analyzed in the paper. The critical values and boundary are obtained. When $\overline{g}_{\mathrm{Na}}$ increases to the critical value, the model will have bifurcation phenomenon, which means system will reach stable state when $\overline{g}_{\mathrm{Na}}$ is less than the critical value and the cell will have continuous action potential after stimulation when $\overline{g}_{\mathrm{Na}}$ is greater. However, there are two critical values about $\overline{g}_{\mathrm{K}}$. The system will be stable when $\overline{g}_{\mathrm{K}}$ is less than the smaller critical value and there are periodic solutions when $\overline{g}_{\mathrm{K}}$ is greater than the value and meanwhile is less than the larger one. The model will reach steady state again when $\overline{g}_{\mathrm{K}}$ is greater than the larger critical value. From analyzing $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ collectively, we can get a critical line which divides the $\overline{g}_{\mathrm{Na}}$-$\overline{g}_{\mathrm{K}}$ plane into two parts. The system will be stable when $(\overline{g}_{\mathrm{Na}}, \overline{g}_{\mathrm{K}})$ is in the upper half plane and model will have periodic solutions when $(\overline{g}_{\mathrm{Na}}, \overline{g}_{\mathrm{K}})$ is in the lower half.

In our analysis, when $\overline{g}_{\mathrm{Na}}$ or $\overline{g}_{\mathrm{K}}$ are taken as the variable(s), all the other parameters are kept with desired values.

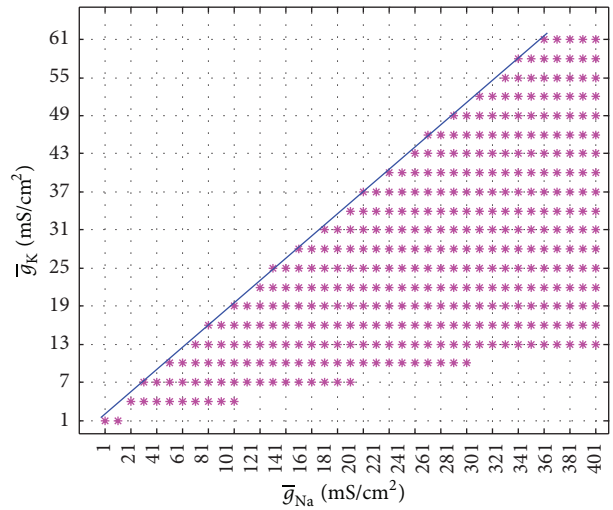


Figure 6: The $\overline{g}_{\mathrm{Na}}$-$\overline{g}_{\mathrm{K}}$ plane and the critical boundary.

However, almost all the biological systems are coupled. All the components influence one another and work together forming the overall functionality. Therefore, when the sodium ($\overline{g}_{\mathrm{Na}}$) and/or potassium ($\overline{g}_{\mathrm{K}}$) channels vary, do the other parameters remain unchanged? Is it reasonable to keep the other parameters still with desired values? We could not ensure that it must be reasonable. Nevertheless, some evidences may explain a certain rationality of the method. For example, tetrodotoxin (TTX) selectively binds to the outer vestibule voltage-gated sodium channels, preventing channels from opening [31]. Ivabradine is a sinus node I_f channel inhibitor, which is selective for the I_f current but does not affect other cardiac ionic currents [32]. Acacetin could suppress the ultrarapid delayed rectifier K^+ current and the transient outward K^+ current and block the acetylcholine-activated K^+ current; however, it has no effect on Na^+ current, L-type Ca^{2+} current, or even inward-rectifier K^+ current [33]. All of these demonstrate that to an extent when one channel changes, the others may not be affected. That is, when the parameter describing a channel varies, it is reasonable to keep parameters describing the other channels unchanged.

Stable states indicate that the electrophysiological activity of cell will get to corresponding resting state at last, while periodic phenomenon looks like response of pathological cell's action potentials caused by cardiac arrhythmias [34]. In other words, $\overline{g}_{\mathrm{Na}}$ and $\overline{g}_{\mathrm{K}}$ may be the causes of the similar diseases to cardiac arrhythmias. So given appropriate medicine to change $\overline{g}_{\mathrm{Na}}$ or $\overline{g}_{\mathrm{K}}$ to reasonable intervals, the corresponding diseases could be abolished or the discomfort can be ameliorated. After all, our research could be a reference to treat relevant diseases. Some diseases led to by abnormal ion channels may be eased by medicine to adjust the conductance into corresponding intervals.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant no. 61173086 and no. 61179009.

References

[1] A. L. Hodgkin and A. F. Huxley, "The components of membrane conductance in the giant axon of *Loligo*," *The Journal of Physiology*, vol. 116, no. 4, pp. 473–496, 1952.

[2] A. L. Hodgkin and A. F. Huxley, "The components of membrane conductance in the giant axon of Loligo," *The Journal of Physiology*, vol. 116, no. 4, pp. 473–496, 1952.

[3] A. L. Hodgkin and A. F. Huxley, "The dual effect of membrane potential on sodium conductance in the giant axon of Loligo," *The Journal of Physiology*, vol. 116, no. 4, pp. 497–506, 1952.

[4] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve.," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[5] C. Soeller, I. D. Jayasinghe, P. Li, A. V. Holden, and M. B. Cannell, "Three-dimensional high-resolution imaging of cardiac proteins to construct models of intracellular Ca^{2+} signalling in rat ventricular myocytes," *Experimental Physiology*, vol. 94, no. 5, pp. 496–508, 2009.

[6] M. Wussling and G. Szymanski, "Simulation by two calcium store models of myocardial dynamic properties: potentiation, staircase, and biphasic tension development," *General Physiology and Biophysics*, vol. 5, no. 2, pp. 135–152, 1986.

[7] A. Mahajan, Y. Shiferaw, D. Sato et al., "A rabbit ventricular action potential model replicating cardiac dynamics at rapid heart rates," *Biophysical Journal*, vol. 94, no. 2, pp. 392–410, 2008.

[8] D. Noble, "Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations," *Nature*, vol. 188, no. 4749, pp. 495–497, 1960.

[9] J. Kneller, R. J. Ramirez, D. Chartier, M. Courtemanche, and S. Nattel, "Time-dependent transients in an ionically based mathematical model of the canine atrial action potential," *American Journal of Physiology—Heart and Circulatory Physiology*, vol. 282, no. 4, pp. H1437–H1451, 2002.

[10] H. Zhang, A. V. Holden, I. Kodama et al., "Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node," *American Journal of Physiology: Heart and Circulatory Physiology*, vol. 279, no. 1, pp. H397–H421, 2000.

[11] A. O. Verkerk and R. Wilders, "Hyperpolarization-activated current, If, in mathematical models of rabbit sinoatrial node pacemaker cells," *BioMed Research International*, vol. 2013, Article ID 872454, 18 pages, 2013.

[12] T. D. Butters, O. V. Aslanidi, J. Zhao, B. Smaill, and H. Zhang, "A novel computational sheep atria model for the study of atrial fibrillation," *Interface Focus*, vol. 3, no. 2, Article ID 20120067, 2013.

[13] T. O'Hara, L. Virág, A. Varró, and Y. Rudy, "Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation," *PLoS Computational Biology*, vol. 7, no. 5, Article ID e1002061, 2011.

[14] P. Li and Y. Rudy, "A model of canine purkinje cell electrophysiology and Ca^{2+} cycling: rate dependence, triggered activity, and comparison to ventricular myocytes," *Circulation Research*, vol. 109, no. 1, pp. 71–79, 2011.

[15] B. Hassard, "Bifurcation of periodic solutions of the Hodgkin-Huxley model for the squid giant axon," *Journal of Theoretical Biology*, vol. 71, no. 3, pp. 401–420, 1978.

[16] J. Rinzel and R. N. Miller, "Numerical calculation of stable and unstable periodic solutions to the Hodgkin-Huxley equations," *Mathematical Biosciences*, vol. 49, no. 1, pp. 27–59, 1980.

[17] K. Aihara and G. Matsumoto, "Two stable steady states in the Hodgkin-Huxley axons," *Biophysical Journal*, vol. 41, no. 1, pp. 87–89, 1983.

[18] J. Guckenheimer and J. S. Labouriau, "Bifurcation of the Hodgkin and Huxley equations: a new twist," *Bulletin of Mathematical Biology*, vol. 55, no. 5, pp. 937–952, 1993.

[19] Y. A. Bedrov, G. N. Akoev, and O. E. Dick, "On the relationship between the number of negative slope regions in the voltage-current curve of the Hodgkin-Huxley model and its parameter values," *Biological Cybernetics*, vol. 73, no. 2, pp. 149–154, 1995.

[20] Y. A. Bedrov, G. N. Akoev, and O. E. Dick, "Partition of the Hodgkin-Huxley type model parameter space into the regions of qualitatively different solutions," *Biological Cybernetics*, vol. 66, no. 5, pp. 413–418, 1992.

[21] H. Fukai, S. Doi, T. Nomura, and S. Sato, "Hopf bifurcations in multiple-parameter space of the Hodgkin-Huxley equations I. Global organization of bistable periodic solutions," *Biological Cybernetics*, vol. 82, no. 3, pp. 215–222, 2000.

[22] K. Terada, H. A. Tanaka, and S. Yoshizawa, "Two-parameter bifurcations in the Hodgkin–Huxley equations for muscle fibers," *Electronics and Communications in Japan*, vol. 83, no. 6, pp. 86–94, 2000.

[23] W. Jiang, C. Yanqiu, F. Xiangyang, and L. Li, "Multi-parameter Hopf-bifurcation in Hodgkin-Huxley model exposed to ELF external electric field," *Chaos, Solitons & Fractals*, vol. 26, no. 4, pp. 1221–1229, 2005.

[24] J. Wang, L. Chen, and X. Fei, "Analysis and control of the bifurcation of Hodgkin-Huxley model," *Chaos, Solitons and Fractals*, vol. 31, no. 1, pp. 247–256, 2007.

[25] J. Wang, L. Q. Chen, and X. Y. Fei, "Bifurcation control of the Hodgkin-Huxley equations," *Chaos, Solitons & Fractals*, vol. 33, no. 1, pp. 217–224, 2007.

[26] Y. Che, J. Wang, B. Deng, X. Wei, and C. Han, "Bifurcations in the Hodgkin-Huxley model exposed to DC electric fields," *Neurocomputing*, vol. 81, pp. 41–48, 2012.

[27] P. M. Hao, D. J. Fan, J. J. Wei, and Q. Liu, "Dynamic behaviors of a delayed HIV model with stage-structure," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 12, pp. 4753–4766, 2012.

[28] X. Y. Chang and J. J. Wei, "Hopf bifurcation and optimal control in a diffusive predator-prey system with time delay and prey harvesting," *Nonlinear Analysis: Modelling and Control*, vol. 17, no. 4, pp. 379–409, 2012.

[29] X. Y. Chang and J. J. Wei, "Stability and Hopf bifurcation in a diffusive predator-prey system incorporating a prey refuge," *Mathematical Biosciences and Engineering*, vol. 10, no. 4, pp. 979–996, 2013.

[30] D. M. Durand and M. Bikson, "Suppression and control of epileptiform activity by electrical stimulation: a review," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1065–1082, 2001.

[31] B. Venkatesh, S. Q. Lu, N. Dandona, S. L. See, S. Brenner, and T. W. Soong, "Genetic basis of tetrodotoxin resistance in pufferfishes," *Current Biology*, vol. 15, no. 22, pp. 2069–2072, 2005.

[32] S. Sulfi and A. D. Timmis, "Ivabradine—the first selective sinus node If channel inhibitor in the treatment of stable angina,"

International Journal of Clinical Practice, vol. 60, no. 2, pp. 222–228, 2006.

[33] G. Li, H. Wang, G. Qin et al., "Acacetin, a natural flavone, selectively inhibits human atrial repolarization potassium currents and prevents atrial fibrillation in dogs," *Circulation*, vol. 117, no. 19, pp. 2449–2457, 2008.

[34] H. O. Wang, D. Chen, and L. G. Bushnell, "Control of bifurcations and chaos in heart rhythms," in *Proceedings of the 36th IEEE Conference on Decision and Control*, vol. 1, pp. 395–400, San Diego, Calif, USA, December 1997.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 386974, 8 pages
http://dx.doi.org/10.1155/2014/386974

Research Article

A Pipeline for Neuron Reconstruction Based on Spatial Sliding Volume Filter Seeding

Dong Sui,[1] Kuanquan Wang,[1] Jinseok Chae,[2] Yue Zhang,[1] and Henggui Zhang[1,3]

[1] *Biocomputing Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[2] *Department of Computer Science and Engineering, Incheon National University, Incheon 402-751, Republic of Korea*
[3] *School of Physics & Astronomy, University of Manchester, Manchester M13 9PL, UK*

Correspondence should be addressed to Kuanquan Wang; wangkq@hit.edu.cn

Received 8 February 2014; Accepted 16 June 2014; Published 2 July 2014

Academic Editor: Huiru Zheng

Neuron's shape and dendritic architecture are important for biosignal transduction in neuron networks. And the anatomy architecture reconstruction of neuron cell is one of the foremost challenges and important issues in neuroscience. Accurate reconstruction results can facilitate the subsequent neuron system simulation. With the development of confocal microscopy technology, researchers can scan neurons at submicron resolution for experiments. These make the reconstruction of complex dendritic trees become more feasible; however, it is still a tedious, time consuming, and labor intensity task. For decades, computer aided methods have been playing an important role in this task, but none of the prevalent algorithms can reconstruct full anatomy structure automatically. All of these make it essential for developing new method for reconstruction. This paper proposes a pipeline with a novel seeding method for reconstructing neuron structures from 3D microscopy images stacks. The pipeline is initialized with a set of seeds detected by sliding volume filter (SVF), and then the open curve snake is applied to the detected seeds for reconstructing the full structure of neuron cells. The experimental results demonstrate that the proposed pipeline exhibits excellent performance in terms of accuracy compared with traditional method, which is clearly a benefit for 3D neuron detection and reconstruction.

1. Introduction

Higher-order cognitive functions in anthropic brain are intricately linked with the processes of nervous system at different biological levels (such as molecular level, cellular level, and system level). The morphological properties of axonal and dendritic arborizations are important aspects of neuronal phenotype. These properties assure the connectivity in the neuron network, thereby facilitate the biological signals transduction in nervous system [1]. Therefore, depicting the function and anatomy structure of neuron cell and networks is of great importance for understanding the way brain works in modern neuron science [2]. Furthermore, great understanding of the mechanism of nervous system can also promote drugs and therapies researching for neurological and psychiatric disease treating.

Extracting neuron morphology from microscopic image data sets is a key point in neurology research. Accurate and efficient reconstruction protocol can facilitate the researches on the function and anatomy structure of neuronal cells and networks. Unfortunately, manually reconstructing neuron structure from microscopy image data sets is labor intensity and time consuming, since the axonal arbors and dendritic are so complex in scale and structure. Therefore, developing new computational methods for neuronal anatomy studying is of particular importance in this context. During the past decades, lots of algorithms and software have been proposed for this task, but most of them achieved limited success.

Since Cohen's team proposed the first fully automated 3D neuron tracing algorithm [3], a large number of approaches have been published for handling the same task in the literature. Generally speaking, these methods can be mainly categorized as minimal path based tracing methods [4, 5], minimum spanning tree methods [6, 7], sequential tracing methods [8, 9], skeletonization methods [3, 10], neuromuscular projection fibers tracing methods [11–16], and active

contour based tracing methods [17, 18]. In the minimal path based tracing methods, algorithms were performed in image subregions instead of the entire image, but these methods cannot extract the exact centerline of tubular structures, such as vessels and neuron fibers. In the minimum spanning tree methods, serious of critical seed points were detected firstly and then the detected seeds were linked into tree representation, such as MDL-MST method and k-MST method [6, 7]. The sequential tracing methods were starting from a set of seed points, but the results of these methods were affected by foreground discontinuity, such as gaps and holes, and these defects required additional post- or preprocessing procedures to overcome [8, 9]. The skeletonization methods mainly relayed on a point-spread function based protocol to trace the neuron anatomy structure, but they were also prone to produce loops and spurs which needed extra postprocessing to smooth the noise [10]. Active contour based methods were particularly attractive for neuron tracing and reconstruction was the most employed protocol for this task [17, 18]. Schmitt's group proposed the first active contour tracing method [17], in which the neuron skeleton was parameterized into a 4D snaxels sets that was characterized by its location and radius [17]. But this method needed to manually set some branching, ending, and other critical points. Vasilkoski and Stepanyants [18] proposed a new method for optimizing the tracing based on the active contours. Following that, Roysam's team proposed an open curve snake based tracing method which was broadly applied in this area [19], and it can allow fully automated processing and user control tracing, but this method can only handle distinct edge neuron images data sets and cannot get accurate neuron radius in vague boundary [19]. Beyond that, there still are some automated tracing tools such as Neuromantic [20], Simple Neurite Tracer [21], NeuronJ [22], and a complete list of the tracing tools that can be founded in the survey paper of Meijering [23]. However, most of these tools still need manual assistance to reconstruct the dendritic and axonal arbors. Therefore, automated 3D neurons anatomy tracing tools need a continuous improvement in the future time.

Traditionally speaking, the pipeline of tracing was initialized by a serious of preprocessing methods, followed by a critical point detection procedure, which was called seeding. Then, these points were linked by center line extraction method, and, finally, radius estimation was applied to reconstruct the full structure [17]. As depicted in Figure 1, the full pipeline was organized in the work flow. In this pipeline, as a key step, excellent seeding method can assure the accuracy of the following skeletonization. There are two approaches for seeding: (i) segmentation and (ii) filtering. The first one is based on a segmentation process, in which the image volumes covered by the neuron were separated from tissue, such as three-dimensional thinning algorithm, but this method is sensitive to noise. The second approach is using a filter to enhance the line elements. Sato's [24] group proposed a 3D multiscale liner image filter to extract the critical property in medical images. This method employed a combination of eigenvalues in hessian matrix of image intensities. Following this approach, Pizer's group proposed another method based on the concept of cores that detected medial points of the object by correlating opposite boundary points [25]. But most of the seed points detected by these methods were distributed unevenly and located at noncritical position [19].

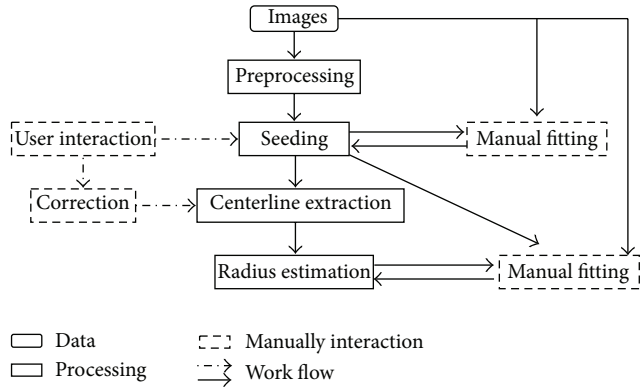


Figure 1: Work flow of neuron tracing pipeline.

Radius estimation is another important part in this pipeline, for it is essential for neuron system simulation [23]. Pock's method was greatly accepted in many tracing tasks and was also used in this paper for the tube-like radius estimation.

In this paper, we proposed a new 3D image filter called sliding volume filter (SVF) to enhance the 3D neuron image data sets and then the most listed voxels were chosen as the final seed points. Then, an open curve snake was employed to reconstruct the neuron anatomy structure. Compared with traditional seed detection method, the SVF method could improve the accuracy of neuron anatomy structure in 3D tracing. Finally, radius estimation was applied to the trace the result for the future functional simulation. And the rest of this paper was organized as follows: data sets collection and method design were illustrated in Section 2, the experimental results and discussions were presented in Section 3, and finally the conclusions were drawn in Section 4.

2. Methodology

Our works were greatly related to Roysam's pipeline for neuron reconstruction, which was based on open curve snake tracing [19]. In this paper, a SVF was designed for seeding by enhancing the spatial tube-like regions and it could provide seed points for the automatic initialization of open-snake models. At last, Pock's method was applied for radius estimation [23].

SVF was expanded from 2D sliding band filter (SBF). As it was depicted in Quelhas' work [26], the 2D SBF could detect rounded convex region in images. It was firstly introduced for detecting cell center in 2D microscopic images [26]. Recently, our research group employed the SBF to detect cells in section images of cat retinal [27] and another transformed SBF to detect insect cells in light field microscopic images [28]. In 3D volume data sets, a rounded convex region was the same as they were in 2D images in gradient vector distribution and we called it spatial convex region.

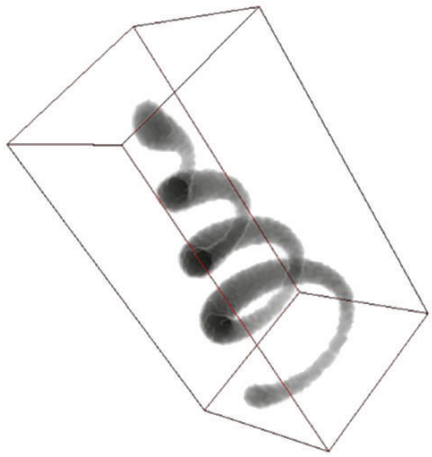


Figure 2: Test data: Helix tube.

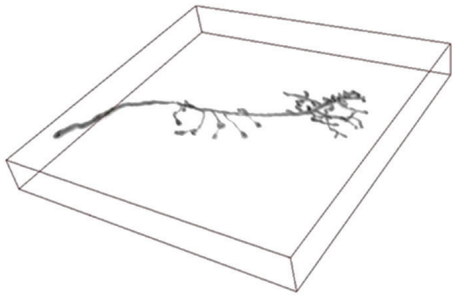


Figure 3: Drosophila olfactory axonal data.

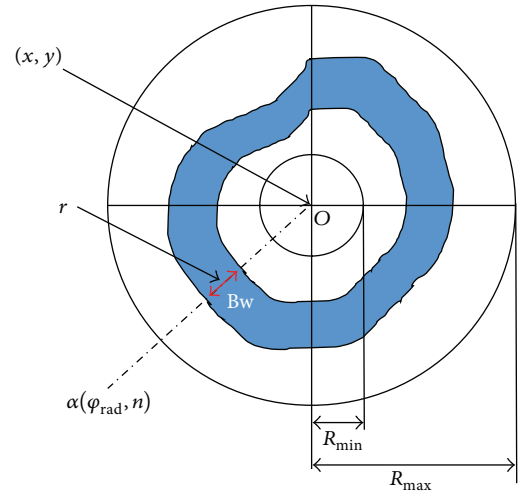


Figure 4: 2D sliding band filter (SBF).

2.1. Data Sets Used in This Paper. In this paper, we choose two kinds of data sets to validate our proposed seeding method. Figure 2 shows a helix image volume data which is a classical test data in neuron tracing [19]. Figure 3 is drosophila olfactory axonal image volume data, and this image data set is firstly designed for single cell label and image registration. Both of these data sets were visualized using Ray casting algorithm in our work and all of the tracing algorithms were performed on these volume data sets.

2.2. 2D Sliding Band Filter (SBF). To introduce SVF, a concept of 2D SBF is important for understanding. The 2D SBF is a member of Convergence Index (CI) family and firstly introduced for detecting cell center in 2D microscopic images [26]. Unlike most of the liner filters' small support regions ($m \times m$ pixels, where $m \in \{2, 3, 5 \ldots\}$), the SBF filter has a larger support region. It has a band with fixed width support region, whose position changes in each radius direction and that allows the maximization of the average Convergence Index in the band width. Figure 4 indicates the support region in SBF and is defined as

$$\mathrm{SBF}(x, y) = \frac{1}{N} \sum_{\mathrm{rad}=1}^{N} \max_{R_{\min} < r < R_{\max}} \left(\frac{1}{\mathrm{Bw}+1} \sum_{r-(\mathrm{Bw}/2)}^{r+(\mathrm{Bw}/2)} \mathrm{CI}(\mathrm{rad}, n) \right), \quad (1)$$

where

$$\mathrm{CI}(\mathrm{rad}, n) = \cos\left(\varphi_{\mathrm{rad}} - \alpha\left(\varphi_{\mathrm{rad}}, n\right)\right),$$
$$\varphi_{\mathrm{rad}} = \frac{2\pi(\mathrm{rad}-1)}{N}, \quad (2)$$
$$\alpha\left(\varphi_{\mathrm{rad}}, n\right) = \arctan\left(\frac{\mathrm{Grad}_{nC}}{\mathrm{Grad}_{nR}}\right),$$

where Grad_{nC} and Grad_{nR} represent the column and row gradient at image position n, N represents the number of support region lines irradiate from the center pixel (x, y), Bw represents the sliding band width, r represents the poison of band center in the support region line ranging from $R_{\min}$ to $R_{\max}$, and $\cos(\varphi_{\mathrm{rad}} - \alpha(\varphi_{\mathrm{rad}}, n))$ represents the angle between the gradient vector at $(\varphi_{\mathrm{rad}}, n)$ and the direction of φ_{rad}.

2.3. SVF Seed Detecting. Before the SVF, this part firstly introduces a concept of Spatial Convergence Index (SCI); see Figure 5. Point $O(x, y, z)$ is the origin in 3D space and the center of support region R. Point p is the voxel in support region R, and its coordinate relative to O is (i, j, k). The radius of the support region R is r. ϕ represents the angle between the gradient vector of P and PO. And the SCI of P relative O is defined as follows:

$$\mathrm{SCI}_{PO}(i, j, k) = \cos\varphi(i, j, k). \quad (3)$$

Then, the SCI of point O in the support region is calculated as

$$\mathrm{SCI}_O = \frac{1}{N} \sum_{p \in R} \cos\varphi\left(i_p, j_p, k_p\right), \quad (4)$$

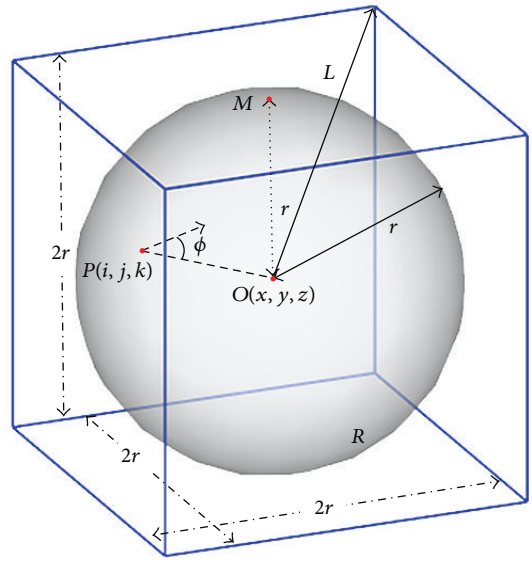


Figure 5: Scheme of spatial convergence index (SCI).

where N is the number of voxel in the support region R. Based on these concepts, the SVF is defined as

$$\begin{aligned}&\mathrm{SVF}(x,y,z)\\&=\frac{1}{M*P_n}\\&\times\sum_{s=0}^{M-1}\sum_{\mathrm{rad}=1}^{P_n}\max_{R_{\min}<r<R_{\max}}\left(\frac{1}{V_t+1}\right.\\&\left.\times\sum_{r-(V_t/2)}^{r+(V_t/2)}\mathrm{SCI}_{\mathrm{OQ}}^{s}(qx_r,qy_r,qz_r)\right),\end{aligned}\tag{5}$$

where

$$\mathrm{SCI}(x,y,z)=\frac{1}{M*P_n}\sum_{s=0}^{M-1}\sum_{(i_m,j_m,k_m)\in R_s}\cos\varphi(i_m,j_m,k_m),\tag{6}$$

where M represents the section in the support region R, P_n represents the support region line in the sth section, and V_t represents the thickness of the sliding volume. The scheme of SVF is depicted in Figure 6.

Seed detection is a critical step before the tracing, it can provide seed points for automatic initialization of the open-snake models. In this paper, seed points are detected by SVF filter voxel by voxel from the start position to the end position in the volume data firstly, and then candidate seeds are chosen if they are extreme in the normal plane of the vessel/axon. Detected seeds are then sorted by the SVF response values, from the largest to the smallest value, and created a seed list for tracing.

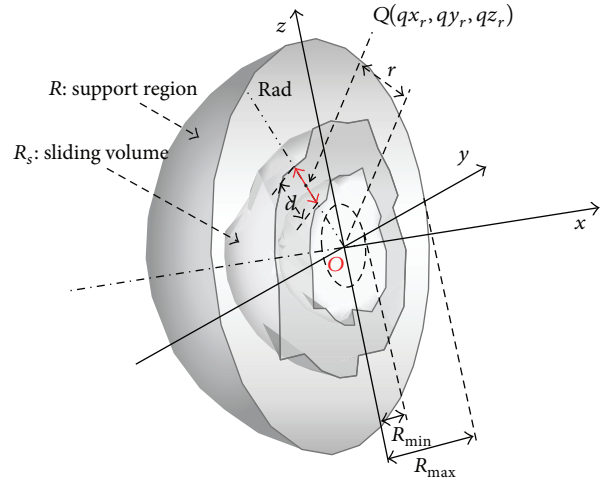


Figure 6: Scheme of sliding volume filter (SVF).

2.4. Neuron Tracing Model. Accurate neuron anatomy structure reconstruction is an import task in neurology. In this part, after initial points selection by the SVF filter, an open-curve snake model for neuron 3D tracing is used for reconstructing the full structure. The open-curve snake is a parametric open curve model. Let $c(s) = (x(s), y(s), z(s))$, $s \in [0, 1]$ and let the snake energy to be minimized as

$$E_{\mathrm{total}}=\int_0^1 E_{\mathrm{int}}(c(s))+E_{\mathrm{ext}}(c(s))\,ds,\tag{7}$$

where $E_{\mathrm{int}}(c(s))$ represents the internal energy for smoothness constraint:

$$E_{\mathrm{int}}(c(s))=\alpha(s)\left|c_s(s)\right|^2+\beta(s)\left|c_{ss}(s)\right|^2,\tag{8}$$

where $\alpha(s)$ and $\beta(s)$ represent the "elasticity" and "stiffness" in the snake, respectively, and

$$E_{\mathrm{External}}=E_{\mathrm{im}}(c(s))+k(s)\,gE_{\mathrm{str}}(c(s)),\tag{9}$$

where

$$\begin{aligned}&\nabla E_{\mathrm{im}}=-\nabla\bar{I}_{\mathrm{GVF}}(x(s),y(s),z(s)),\\&\nabla E_{\mathrm{str}}(c(s))\\&=-\begin{cases}\mathrm{sign}\left(-\dfrac{c_s(s)}{\|c_s(s)\|}\cdot ev_1(c(s))\right)\cdot ev_1(c(s)) & s=0\\ \mathrm{sign}\left(-\dfrac{c_s(s)}{\|c_s(s)\|}\cdot ev_1(c(s))\right)\cdot ev_1(c(s)) & s=1\\ 0 & s\in(0,1).\end{cases}\end{aligned}\tag{10}$$

In (8), $\alpha(s)$ and $\beta(s)$ are "elasticity coefficient" and "stiffness coefficient," respectively, in internal energy, and they embedded the regularity of the curve. $\alpha(s)$ was selected to be 0 for $s \in [0, 1]$, and set $\beta(s)$ was selected to be 0 at $s = 0$ and $s = 1$. In (10), the external energy term is employed for making the snake deform along the center line of the neuron fiber, where

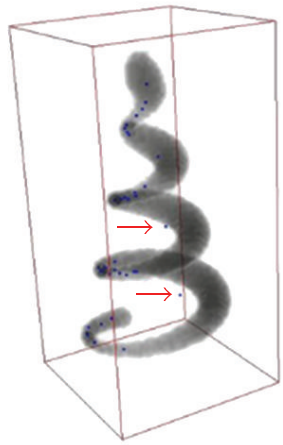


Figure 7: Thresholding seeds detection method.

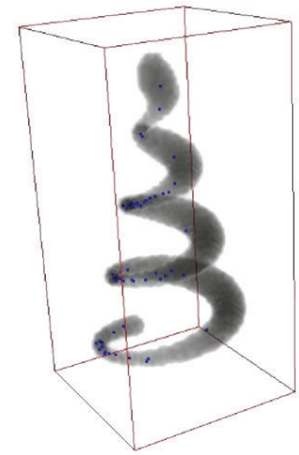


Figure 8: SVF seeds detection method.

∇E_{im} is the negative normalized gradient vector flow, $k(s)$ is a weighted parameter, $ev_1(c(s))$ is the first principal direction of the Jacobian matrix, and the $\nabla E_{\text{str}}(c(s))$ is nonzero item when it is located at the tail and the end pointing to the right direction of the snake.

2.5. Radius Estimation. In the following work, to provide more detailed information for functional simulation, the radius estimation is applied to each point on the snake after tracing. As shown in Figure 9, Pock's method is applied for boundaries measurement to detect the tube edge [23]. The edge of the tube-like volume is B, and it is defined as a circular centered at O which is a seed point, shown in Figure 9. Equation (11) describes the boundary as follows:

$$B(o,r) = \frac{1}{N}\sum_{i=1}^{N} \text{grad}(o + rv_{\text{ai}})\, \text{gmax}\left(-\nabla \bar{I}_{\text{GVF}}(o + rv_{\text{ai}})\, \text{g}v_{\text{ai}}, 0\right), \tag{11}$$

where $v_{\text{ai}} = \cos(\text{ai})v_1 + \sin(\text{ai})v_2$ is the radial vector in the v_1-v_2 plane of point O on the snake and $\text{grad}(o + rv_{\text{ai}})$ is the gradient magnitude of the point on the circle. The radius r is sampled in the circle by a certain angle distance and in this paper the N is set as 8 in the radius circle.

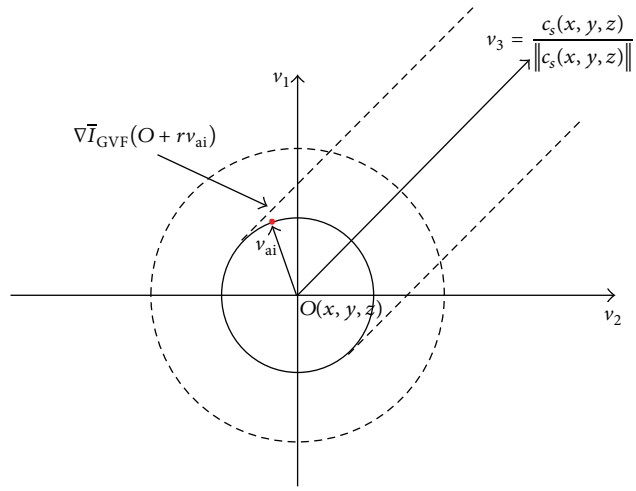


Figure 9: Illustration of the circular cross-section. v_3 is the tangent vector in point O, and v_1 and v_2 are the two orthogonal vectors defining the normal plane.

3. Results and Discussion

For all the image data sets, the following parameters were chosen as the default setting by visual estimation of diameter in average radius of the neuron cross-section. For seeds detection, parameter $M = 32$, $P_n = 32$ was chosen, respectively, the remaining parameters of SVF were chosen as $V_t = 8$, $R_{\text{min}} = 10$, and $R_{\text{max}} = 30$, and the unit of all parameters was "pixels." Comparison of test data sets seeding results between threshold method and SVF method results has shown an excellent detection results of SVF method in seed detection. Figure 7 shows a traditional threshold seeding method and there are some seed points that fall out of the edge of real Helix body, which are headlight with red arrows. And Figure 8 is of the same perspective as Figure 7 and shows that the SVF seeding method can detect most of the critical points as candidate seeds. After tracing from the detecting seeds points by SVF, the whole structure of test data sets is generated accurately which are shown in Figure 10.

When it is applied to real data sets, the SVF seeding method can detect most of the critical seed points in the body of olfactory axonal, shown in Figure 11. Threshold seeding method is not shown here for its poor results. After specifying the branching points, the tracing result of the open curve snake is shown in Figure 12, and it clearly indicates that nearly all of the anatomy structure is reconstructed after tracing.

After tracing the full structure, the radius of the olfactory axonal is estimated as a following-up procedure for functional simulation. As shown in Figure 13, the gray and black area represents the body of the olfactory axonal, the green line represents the central line of the olfactory axonal, and the

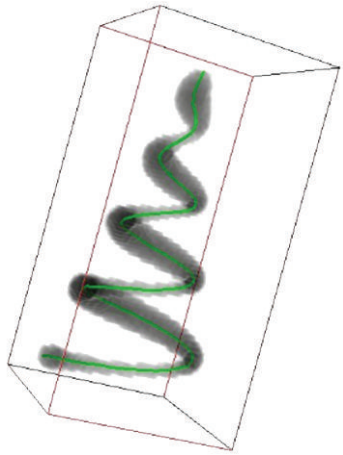


Figure 10: Tracing results by SVF seeds.

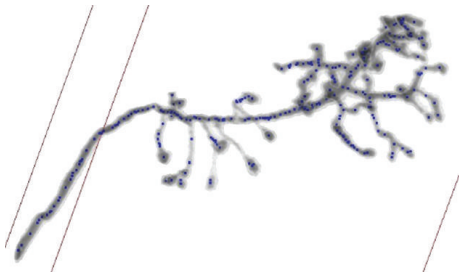


Figure 11: SVF seed detection results for olfactory axonal.

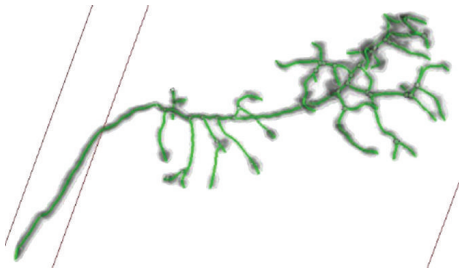


Figure 12: Open curve snake tracing results of olfactory axonal.

blue ring represents the radius of each part from the central line, and it depicts that most of the radius is estimated by the method. Figure 14 exhibits a magnification of the red rectangle area in Figure 13. And another magnification of red rectangle area in Figure 14 is shown in Figure 15.

In Figure 15, the width of olfactory axonal, estimated radius, and center line are marked separately. This task is for the future functional neuronal simulation which is not discussed in this paper.

4. Conclusion

In this paper, a novel seeding method based on spatial SVF is proposed for neuron reconstruction from microscopic image

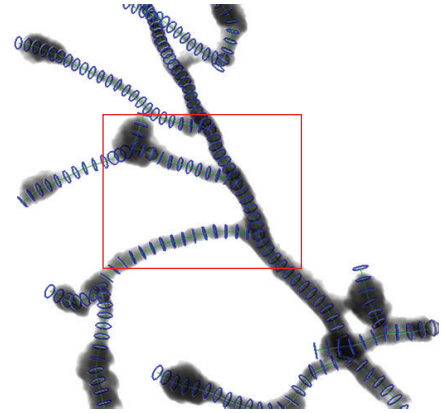


Figure 13: Radius estimation of olfactory axonal.

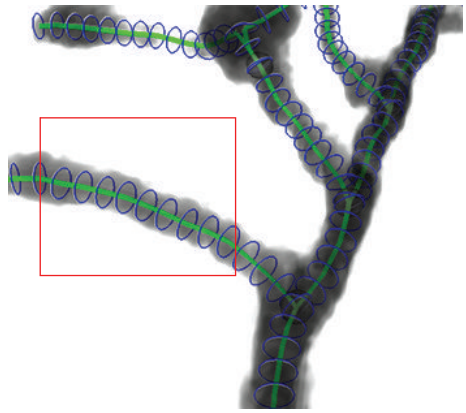


Figure 14: Magnification part of olfactory axonal radius estimation in Figure 11.

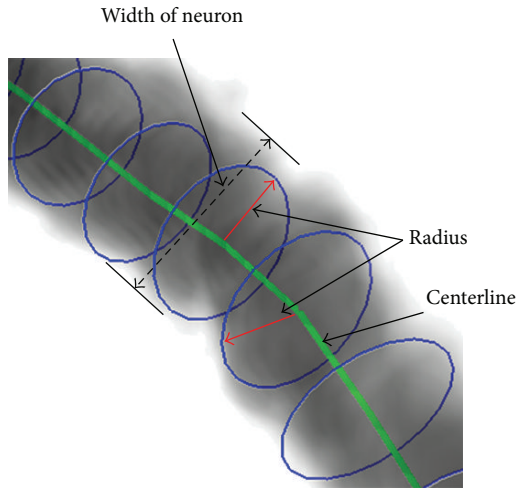


Figure 15: Magnification part of olfactory axonal radius estimation in Figure 12 and detail of radius estimation.

data sets which were collected by confocal microscopy. The seeding results comparison shows that the SVF method can detect seed points accurately in test data sets and detect most of the critical points in olfactory axonal data sets. After open curve snake tracing, both of the data set's structures are reconstructed from SVF seeds. In the last part of our work, a radius estimation method is applied to the tracing result for future functional simulation.

Finally, it is worth noting that this method can clearly be a benefit for seeding task in the protocol of neuron tracing. However, uneven illumination produced by a microscope is also a critical factor influenced the seeding accuracy. Therefore, some illumination correction methods will be studied to improve our method in the future works.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partly supported by National Nature Science Foundation of China (NSFC) Grant no. 61173086 and the University of Incheon International Cooperative Research Grant in 2012.

References

[1] G. A. Ascoli, "Neuroinformatics grand challenges," *Neuroinformatics*, vol. 6, no. 1, pp. 1–3, 2008.

[2] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience, Exploring the Brain*, Lippincott Williams & Wilkins, Baltimore, Md, USA, 3rd edition, 2007.

[3] A. R. Cohen, B. Roysam, and J. N. Turner, "Automated tracing and volume measurements of neurons from 3-D confocal fluorescence microscopy data," *Journal of Microscopy*, vol. 173, part 2, pp. 103–114, 1994.

[4] E. Meijering, M. Jacob, J.-C. F. Sarria, P. Steiner, H. Hirling, and M. Unser, "Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images," *Cytometry A*, vol. 58, no. 2, pp. 167–176, 2004.

[5] H. Peng, Z. Ruan, D. Atasoy, and S. Sternson, "Automatic reconstruction of 3D neuron structures using a graph-augmented deformable model," *Bioinformatics*, vol. 26, no. 12, pp. i38–i46, 2010.

[6] X. Yuan, J. T. Trachtenberg, S. M. Potter, and B. Roysam, "MDL constrained 3-d grayscale skeletonization algorithm for automated extraction of dendrites and spines from fluorescence confocal images," *Neuroinformatics*, vol. 7, no. 4, pp. 213–232, 2009.

[7] G. González, E. Türetken, F. Fleuret, and P. Fua, "Delineating trees in noisy 2D images and 3D image-stacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2799–2806, San Francisco, Calif, USA, June 2010.

[8] K. A. Al-Kofahi, S. Lasek, D. H. Szarowski et al., "Rapid automated three-dimensional tracing of neurons from confocal image stacks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 2, pp. 171–187, 2002.

[9] S. R. Aylward and E. Bullitt, "Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 61–75, 2002.

[10] W. He, T. A. Hamilton, A. R. Cohen et al., "Automated three-dimensional tracing of neurons in confocal and brightfield images," *Microscopy and Microanalysis*, vol. 9, no. 4, pp. 296–310, 2003.

[11] H. Cai, X. Xu, J. Lu, J. W. Lichtman, S. P. Yung, and S. T. C. Wong, "Repulsive force based snake model to segment and track neuronal axons in 3D microscopy image stacks," *NeuroImage*, vol. 32, no. 4, pp. 1608–1620, 2006.

[12] H. Cai, X. Xu, J. Lu, J. Lichtman, S. P. Yung, and S. T. C. Wong, "Using nonlinear diffusion and mean shift to detect and connect cross-sections of axons in 3D optical microscopy images," *Medical Image Analysis*, vol. 12, no. 6, pp. 666–675, 2008.

[13] J. Lu, J. C. Fiala, and J. W. Lichtman, "Semi-automated reconstruction of neural processes from large numbers of fluorescence images," *PLoS ONE*, vol. 4, no. 5, Article ID e5655, 2009.

[14] R. Srinivasan, X. Zhou, E. L. Miller, J. Lu, J. W. Lichtman, and S. T. C. Wong, "Automated axon tracking of 3D confocal laser scanning microscopy images using guided probabilistic region merging," *Neuroinformatics*, vol. 7, no. 1, p. 83, 2009.

[15] R. Srinivasan, Q. Li, X. Zhou, J. Lu, J. Lichtman, and S. T. C. Wong, "Reconstruction of the neuromuscular junction connectome," *Bioinformatics*, vol. 26, no. 12, Article ID btq179, pp. i64–i70, 2010.

[16] J. Wang, X. Zhou, J. Lu, J. Lichtman, S. Chang, and S. T. C. Wong, "Dynamic local tracing for 3D axon curvilinear structure detection from microscopic image stack," in *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '07)*, pp. 81–84, Arlington, Va, USA, April 2007.

[17] S. Schmitt, J. F. Evers, C. Duch, M. Scholz, and K. Obermayer, "New methods for the computer-assisted 3-D reconstruction of neurons from confocal image stacks," *NeuroImage*, vol. 23, no. 4, pp. 1283–1298, 2004.

[18] Z. Vasilkoski and A. Stepanyants, "Detection of the optimal neuron traces in confocal microscopy images," *Journal of Neuroscience Methods*, vol. 178, no. 1, pp. 197–204, 2009.

[19] Y. Wang, A. Narayanaswamy, C.-L. Tsai, and B. Roysam, "A broadly applicable 3-D neuron tracing method based on open-curve snake," *Neuroinformatics*, vol. 9, no. 2-3, pp. 193–217, 2011.

[20] Neuromantic: the Freeware Neuronal Reconstruction Tool, http://www.reading.ac.uk/neuromantic.

[21] Simple Neurite Tracer, http://pacific.mpi-cbg.de/wiki/index.php/Simple_Neurite_Tracer.

[22] NeuronJ, http://www.imagescience.org/meijering/software/neuronj/manual.html.

[23] E. Meijering, "Neuron tracing in perspective," *Cytometry A*, vol. 77, no. 7, pp. 693–704, 2010.

[24] Y. Sato, S. Nakajima, N. Shiraga et al., "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," *Medical Image Analysis*, vol. 2, no. 2, pp. 143–168, 1998.

[25] S. M. Pizer, D. Eberly, D. S. Fritsch, and B. S. Morse, "Zoom-invariant figural shape: the mathematics of cores," *Computer Vision and Image Understanding*, vol. 69, no. 1, pp. 55–71, 1998.

[26] P. Quelhas, M. Marcuzzo, A. M. Mendonça, and A. Campilho, "Cell nuclei and cytoplasm joint segmentation using the sliding band filter," *IEEE Transactions on Medical Imaging*, vol. 29, no. 8, pp. 1463–1473, 2010.

[27] D. Sui and K. Wang, "A counting method for density packed cells based on sliding band filter image enhancement," *Journal of Microscopy*, vol. 250, no. 1, pp. 42–49, 2013.

[28] K. Wang, D. Sui, W. Wang, Y. Yuan, and W. Zuo, "A cell counting method for BEVS based on nonlinear transformed sliding band filter," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 12)*, vol. 2012, pp. 118–121, San Diego, Calif, USA, September 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 891842, 9 pages
http://dx.doi.org/10.1155/2014/891842

Research Article

Correlating Information Contents of Gene Ontology Terms to Infer Semantic Similarity of Gene Products

Mingxin Gan

Dongling School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China

Correspondence should be addressed to Mingxin Gan; ganmx@ustb.edu.cn

Received 29 January 2014; Accepted 29 April 2014; Published 22 May 2014

Academic Editor: Huiru Zheng

Successful applications of the gene ontology to the inference of functional relationships between gene products in recent years have raised the need for computational methods to automatically calculate semantic similarity between gene products based on semantic similarity of gene ontology terms. Nevertheless, existing methods, though having been widely used in a variety of applications, may significantly overestimate semantic similarity between genes that are actually not functionally related, thereby yielding misleading results in applications. To overcome this limitation, we propose to represent a gene product as a vector that is composed of information contents of gene ontology terms annotated for the gene product, and we suggest calculating similarity between two gene products as the relatedness of their corresponding vectors using three measures: Pearson's correlation coefficient, cosine similarity, and the Jaccard index. We focus on the biological process domain of the gene ontology and annotations of yeast proteins to study the effectiveness of the proposed measures. Results show that semantic similarity scores calculated using the proposed measures are more consistent with known biological knowledge than those derived using a list of existing methods, suggesting the effectiveness of our method in characterizing functional relationships between gene products.

1. Introduction

Over the last few years, domain ontologies have been successfully applied to describe entities within a variety of biological domains, with examples including the derivation of functional relationships between gene products based on the gene ontology (GO) [1–3], the inference of phenotype similarity between human diseases based on the human phenotype ontology (HPO) [4, 5], the modeling of general computational tasks in systems biology based on the systems biology ontology (SBO) [6], and many others [7–9]. With an ontology to provide controlled and structured vocabularies in a specific biological domain and annotations to characterize entities in the domain with the vocabularies, relationships between the entities can be quantified by their semantic similarities in the ontology, thereby providing a convenient yet powerful means of profiling the entities and their semantic relationships [1]. Nevertheless, the automated derivation of semantic similarity between entities based on their annotations in a domain specific ontology still remains a great challenge, appealing for the development of effective and convenient computational methods [10].

In general, a domain ontology provides a set of controlled and relational vocabularies for describing domain specific knowledge. The vocabularies, also referred to as concepts or terms, are often organized as a directed acyclic graph (DAG), in which vertices denote terms and edges represent semantic relationships between the terms. It is also common that an ontology has more than one semantic relationship. For example, in the gene ontology, there are multiple types of semantic relationships such as "A is_a B" (any instance of A is also an instance of B) and "A part_of B" (an instance of A is a component of some instances of B) [1]. Given such a domain specific ontology and annotations that map entities onto the terms, most existing methods first calculate pairwise semantic similarity between the terms using the structure

of the ontology and annotations of entities and then derive similarity between the entities based on similarity between the terms [10–14].

Taking the gene ontology as an example, in order to achieve the former objective, Resnik proposed to use the information content (the negative logarithm of the relative frequency of occurrence of a term in annotations for a set of gene products) of the lowest common ancestor of two query terms to measure their semantic similarity [11]. Lin modified this measure by taking information contents of the query terms into consideration [12]. Schlicker et al. further incorporated the relative frequency of occurrence of the lowest common ancestor into the measure of Lin [14]. Jiang and Conrath proposed to incorporate the information contents of the query terms by using a formula different from that of Lin [13]. As another branch, Wang et al. proposed to calculate semantic similarity between GO terms using only the structural information of the underlying gene ontology, with the consideration of two types of semantic relationships: is_a and part_of [10].

With similarities between GO terms calculated, the semantic similarity between two query gene products was often calculated using a mean-max rule [10]. More specifically, given a single GO term and a collection of GO terms, the similarity between the term and the collection was defined as the maximum similarity between the term and every term in the collection. Furthermore, the similarity between two collections of GO terms was defined as the average of similarity between every term in a collection and the other collections. Finally, since a gene product was annotated by a collection of GO terms, semantic similarity between two gene products was defined as the similarity between the corresponding two sets of GO terms.

The above methods have been successfully applied to a variety of fields, with examples including the calculation of functional similarity between proteins based on the gene ontology (GO) for the inference of disease genes [2], the characterization of phenotype similarity between human diseases based on the human phenotype ontology (HPO) [5], and many others [7]. Software packages implementing these methods have also been released and publically available in the community of bioinformatics and computational biology, with examples including GOSemSim [15], FuSSiMeG [16], and OWLSim [4]. However, disadvantages of these methods are also obvious. For example, although methods such as those in [12–14] took efforts to modify the method of Resnik [11], their methods often performed worse than that of Resnik in real applications [10], suggesting that the revision of information contents can hardly be effective. Also, although Wang et al. systematically considered the structure and multiple semantic relationships of the gene ontology [10], they discarded the valuable resource of information contents of GO terms, resulting in a method performing worse than that of Resnik in many applications such as the prioritization of candidate genes [2]. In addition, as we shall see in the Results section, all of these methods tend to overestimate similarity between proteins that are actually not similar in their functions, thereby yielding misleading results in applications.

With these understandings, we propose in this paper to represent a gene product using a vector that is composed of information contents of GO terms annotated for the product in the gene ontology. Based on this notion, we suggest calculating semantic similarity between gene products as the relatedness of their corresponding vectors using three measures: Pearson's correlation coefficient, cosine similarity, and the Jaccard index. We focus on the biological process namespace of the gene ontology and annotations of proteins of the budding yeast *Saccharomyces cerevisiae* to perform a series of comprehensive studies on the effectiveness of the proposed measures. We calculate semantic similarity scores between yeast genes relying on the biological process domain of the gene ontology, use the resulting semantic similarity scores to measure functional relationships between the proteins, and study the consistency between such relationships and known biological knowledge. Results on 141 yeast biochemical pathways, 1,022 protein families, and two large-scale yeast protein-protein interaction networks show that semantic similarity scores calculated using the proposed measures are more consistent with biological knowledge than those derived using a list of existing methods, suggesting the effectiveness of our method in characterizing semantic similarity between gene products.

2. Methods

2.1. The Gene Ontology and Species Specific Annotations. The gene ontology (GO) provides a controlled vocabulary of terms for describing characteristics of gene products. This ontology covers three domains: biological process (BP), molecular function (MF), and cellular component (CC). The biological process domain defines operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of living cells, tissues, organs, and organisms. The molecular function domain represents the elemental activities of a gene product at the molecular level, such as binding or catalysis. The cellular component domain describes the parts of a cell or its extracellular environment [1]. Each of these three domains is organized according to a directed acyclic graph (DAG) structure, represented as $G = (V, E)$, where V is a set of vertices denoting concepts and E is a set of edges denoting semantic relationships between the terms. In such a graph, we use P_t and C_t to denote the sets of parents and children of term t, including t itself, respectively, and we use A_t and D_t to denote ancestors and descendants of term t, including t itself, respectively. Note that in the gene ontology, there are multiple types of semantic relationships such as "A is_a B" (any instance of A is also an instance of B) and "A part_of B" (an instance of A is a component of some instance of B).

A species specified annotation provides a mapping from a gene product of the species to a term in a domain (BP, MF, or CC) of the gene ontology. Following common specifications, the annotation of a gene product with term t implies the annotation of the gene product with all ancestors of t. With this notion, we represent annotations of gene product g using a binary annotation vector $\mathbf{a}_g = (a_{gi})_{|V|\times 1}$, where $a_{gi} = 1$ if g

is annotated by the term indexed by i or its descendants and $|V|$ the total number of terms in a domain.

3. Semantic Similarity as Correlation of Information Contents

Given a domain of the gene ontology and annotations for a set of gene products, the probability that a product annotated by term t or its descendants is estimated using the relative frequency of occurrence of term t and its descendants in the annotations is calculated by

$$\Pr(t) = \frac{1}{N}\sum_{i\in D_t} n_i, \tag{1}$$

where n_i is the number of annotations with term i and N the total number of annotations. The information content of term t is then calculated as

$$\mathrm{IC}(t) = -\log \Pr(t). \tag{2}$$

Moreover, information contents of all terms in the domain can be represented as a vector $\mathbf{q} = (q_i)_{|V|\times 1}$ with q_i being the information content of the term indexed by i. Calculating the Hadamard (entrywise) product of $\mathbf{a}_g$ and $\mathbf{q}$, we obtain the vector of information contents for gene product g as $\mathbf{x}_g = \mathbf{q} \circ \mathbf{a}_g = (x_{gi})_{|V|\times 1}$, where $x_{gi} = q_i \times a_{gi}$ for $i = 1, \ldots, |V|$. With such a vector calculated for every gene product, we propose the following three measures to quantify semantic similarity between two entities.

First, we propose to calculate the similarity as the absolute value of Pearson's correlation coefficient between the two vectors $\mathbf{x}_g$ and $\mathbf{x}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{correlation})} = \left| \frac{\sum_{1\le i\le |V|} \left(x_{gi} - \overline{x}_g\right)\left(x_{hi} - \overline{x}_g\right)}{\sqrt{\sum_{1\le i\le |V|} \left(x_{gi} - \overline{x}_g\right)^2}\sqrt{\sum_{1\le i\le |V|} \left(x_{hi} - \overline{x}_h\right)^2}} \right|. \tag{3}$$

In this measure, we assume that information contents for the two gene products, $\mathbf{x}_g$ and $\mathbf{x}_h$, have a linear relationship, say,

$$\mathbf{x}_g = \alpha + \beta \mathbf{x}_h. \tag{4}$$

Hence, it is natural to use the coefficient of determination (r^2) that measures how good the observations fit this linear model to quantify the similarity between the two vectors. To ease the computation, we simply calculate the absolute value of the correlation coefficient instead of r^2. Note that exchanging $\mathbf{x}_g$ and $\mathbf{x}_h$ in the linear model yields the same r^2.

Second, we calculate the similarity as the cosine of the angle between the two vectors $\mathbf{x}_g$ and $\mathbf{x}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{cosine})} = \frac{\sum_{1\le i\le |V|} x_{gi}x_{hi}}{\sqrt{\sum_{1\le i\le |V|} x_{gi}^2}\sqrt{\sum_{1\le i\le |V|} x_{hi}^2}}. \tag{5}$$

This is equivalent to calculating the uncentered correlation coefficient of the two vectors. It is evident that the cosine measure will yield similar results as those of the correlation measure when the means of $\mathbf{x}_g$ and $\mathbf{x}_h$ are small.

Third, we calculate the similarity as the Jaccard index of the two annotation vectors $\mathbf{a}_g$ and $\mathbf{a}_h$ for two gene products g and h as

$$S_{gh}^{(\mathrm{Jaccard})} = \frac{\sum_{1\le i\le |V|} \left(a_{gi} \wedge a_{hi}\right)}{\sum_{1\le i\le |V|} \left(a_{gi} \vee a_{hi}\right)}. \tag{6}$$

This is equivalent to calculating the ratio of the number of elements in the intersection and union of the two annotation sets for gene products g and h.

4. Existing Methods for Calculating Semantic Similarity

Most existing methods first derive similarity scores between terms and then calculate semantic similarity scores between gene products as similarity scores between collections of annotated terms for the products. More precisely, there have been two main categories of methods for calculating pairwise concept similarity scores: (1) approaches based on information contents of terms in the gene ontology and (2) methods based on the structure of the gent ontology.

The first group of approaches calculates similarity between two terms u and v relying on the information content of the most specific term m_{uv} in their common ancestors. Generally, a term with more specific meaning tends to have a higher information content and hence

$$m_{uv} = \arg \max_{w\in A_u\cap A_v} \mathrm{IC}(w). \tag{7}$$

With this notion, Resnik [11] defined the similarity between u and v as

$$T_{uv}^{(\mathrm{Resnik})} = \mathrm{IC}(m_{uv}) = -\log \Pr(m_{uv}). \tag{8}$$

Lin [12] defined the similarity as

$$T_{uv}^{(\mathrm{Lin})} = \frac{2\log \Pr(m_{uv})}{\log \Pr(u) + \log \Pr(u)}. \tag{9}$$

Schlicker et al. [14] define the similarity as

$$T_{uv}^{(\mathrm{Schlicker})} = \frac{2\log \Pr(m_{uv})}{\log \Pr(u) + \log \Pr(u)}\left(1 - \Pr(m_{uv})\right). \tag{10}$$

Jiang and Conrath [13] define the dissimilarity between two terms as

$$D_{uv}^{(\mathrm{Jiang})} = \log \Pr(u) + \log \Pr(u) - 2\log \Pr(m_{uv}). \tag{11}$$

This is equivalent to defining its reciprocal as the similarity as

$$T_{uv}^{(\mathrm{Jiang})} = \frac{1}{\log \Pr(u) + \log \Pr(u) - 2\log \Pr(m_{uv})}. \tag{12}$$

The second group of approaches calculates similarity between GO terms depending on the structure of the gene ontology. Briefly, given a term indexed by t, Wang et al. iteratively calculate an s-value for every ancestor $a \in A_t$ to measure the contribution of a to the semantic of t as

$$s_t(a) = \begin{cases} 1 & \text{if } a = t, \\ \max_{x \in C_a} w_e s_t(x) & \text{if } a \neq t, \end{cases} \quad (13)$$

where the weight $w_e = 0.8$ if x and t have the is_a relationship and $w_e = 0.6$ if x and t have the part_of relationship [10]. Then, a semantic value for term t is defined as $s(t) = \Sigma_{x \in A_t} s_t(x)$. Finally, the semantic similarity score between two terms u and v is defined as

$$T_{uv}^{(\text{Wang})} = \sum_{x \in A_u \cap A_v} \frac{s_u(x) + s_v(x)}{s(u) + s(v)}. \quad (14)$$

With pairwise semantic similarity scores between GO terms being ready, the similarity between term t and a set of terms T is defined as

$$\text{Sim}(t, T) = \max_{t' \in T} T_{tt'}, \quad (15)$$

where $T_{tt'}$ is calculated using either of the above methods. The similarity between two sets of terms S and T can then be calculated as

$$\text{Sim}(S, T) = \frac{1}{|S| + |T|} \left(\sum_{s \in S} \text{Sim}(s, T) + \sum_{t \in T} \text{Sim}(t, S) \right). \quad (16)$$

Finally, for two gene products g and h annotated by two sets of terms G and H, respectively, the semantic similarity between the two objects is then defined as

$$S_{gh} = \text{Sim}(G, H). \quad (17)$$

5. Results

5.1. Data Sources. There have been quite a few domain specific ontologies available for characterizing entities in a variety of biological domains. Particularly, the OBO (open biological and biomedical ontologies) Foundry has released eight ontologies to provide standard descriptions of entities in biological domains [14]. Among these ontologies, biological process (BP), molecular function (MF), and cellular component (CC) are typically referred to as the gene ontology (GO), which has been widely used to describe functions of genes. The gene ontology also provides annotations of gene products for several well-studied model organisms, including yeast, fruit fly, and mouse [1]. In this paper, we focus on the biological process domain of GO and annotations of the budding yeast *Saccharomyces cerevisiae* to validate the effectiveness of the proposed measures. We extract 22,688 terms from the biological process domain of the gene ontology (released on April 27, 2012) and obtain 22,798 annotations of 6,383 yeast genes (released on April 28, 2012).

5.2. Distribution of Semantic Similarity Scores of Random Gene Pairs. It is evident that a pair of genes selected at random can hardly have similar functions, and thus the semantic similarity score between such a pair of genes should be close to zero. To validate this argument, we calculate semantic similarity scores of 100,000 pairs of yeast genes selected at random, and we summarize the distribution of the scores in Figure 1. We can clearly see from the figure that the median similarity score of the correlation measure (0.004894) is almost 0 so is that of the cosine measure (0.003196). The median similarity score of the Jaccard measure (0.03846) is higher than those for both the correlation and the cosine measures but still lower than those for all the five existing methods. The method of Resnik generates the smallest median similarity score (0.04395) among the existing methods, followed by the methods of Schlicker et al. (0.04810), Lin (0.09115), and Wang et al. (0.2138). The method of Jiang et al. generates the largest median similarity score (0.3460). From these observations, we conclude that the existing methods tend to overestimate semantic similarity between genes that are actually not related in their functions. On the other hand, the proposed measures, though much simpler than the existing methods, do not have such a drawback and thus yield much more reasonable results in assessing semantic similarity between randomly selected gene pairs.

5.3. Consistency between Gene Semantic Similarity and Pathway Data. It is known that most biological functions rise from collaborative effects of several proteins that usually involve in the same biological process and form a pathway [17]. Hence, gene products (proteins) in the same pathway should have similar annotations in the biological process ontology and in turn own high semantic similarity scores according this ontology. On the contrary, gene products belonging to different pathways should own relatively low semantic similarity scores. To assess whether the proposed similarity measures are consistent with this knowledge, we compare semantic similarity scores between proteins within a pathway and those between proteins involved in different pathways as follows.

We download from the *Saccharomyces* Genome database (SGD) [18] 141 pathways, each including at least two proteins. For each of these pathways, we calculate pairwise semantic similarity scores of proteins involved in the pathway, and we average these scores over all pairs of proteins to obtain the mean semantic similarity score within the pathway (μ_{in}). Meanwhile, for each pathway, we further select at random 10 times the number of proteins as those in the pathway, calculate semantic similarity scores between these proteins and those in the pathway, and average over these scores to obtain the mean semantic similarity score outside the pathway (μ_{out}). Then, we plot the distribution of mean similarity scores within and outside all pathways in Figure 2. From the figure, we observe that the mean similarity scores within pathways are in general large, while those outside pathways are typically small. Particularly, for all of the three proposed measures (correlation, cosine, and the Jaccard), the differences between the medians of the mean similarity

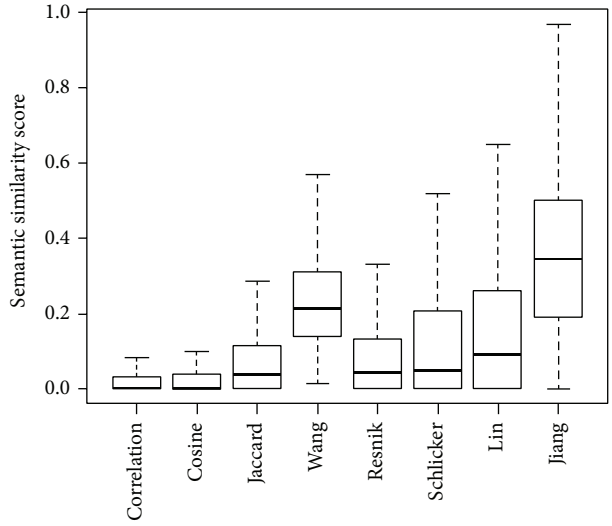


Figure 1: Distributions of semantic similarity scores of 100,000 randomly selected pairs of yeast genes.

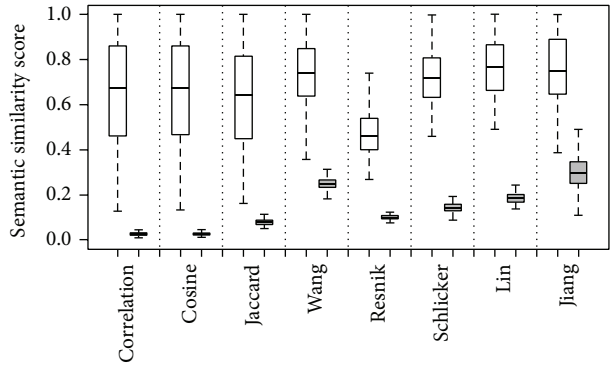


Figure 2: Distributions of mean semantic similarity scores within pathways (white) and outside pathways (gray).

scores within and outside pathways are much more obvious than those of the five existing methods. For example, using the correlation measure, we obtain the median μ_{in} over all pathways as 0.6578 and the median μ_{out} as 0.02564. Using the cosine measure, we obtain a median μ_{in} of 0.6600 and a median μ_{out} of 0.02733. In contrast, the method of Wang produces a median μ_{in} of 0.7405 and a median μ_{out} of 0.2489, and the method of Resnik produces a median μ_{in} of 0.4662 and a median μ_{out} of 0.09956.

We further calculate for each pathway the ratio of the mean semantic similarity scores within the pathway over that outside the pathway ($\mu_{\text{in}}/\mu_{\text{out}}$), and we average such ratios over all 141 pathways to obtain a criterion called fold change of semantic similarity scores within pathways against those outside pathways. We summarize the fold changes in Figure 3, from which we can clearly see the effectiveness of the proposed measures. For example, using the correlation measure, we obtain a fold enhancement of 29.93. Using the cosine measure, we obtain a fold change of 26.65. In contrast, the method of Wang only produces a fold change of 3.03, and

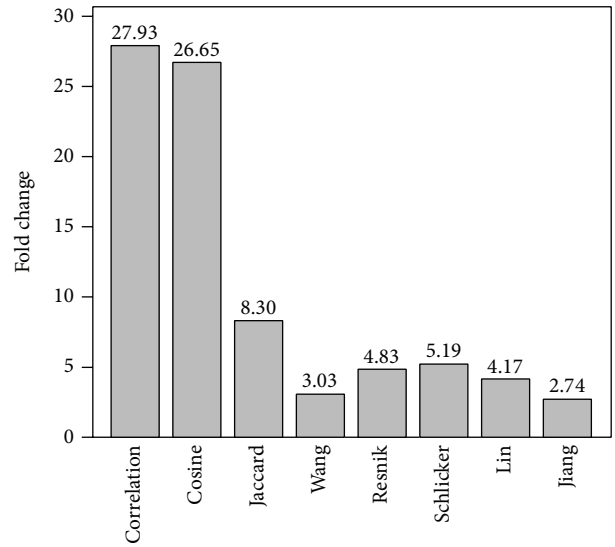


Figure 3: Fold change of semantic similarity scores within pathways against those outside pathways.

the method of Resnik produces a slightly larger fold change of 4.83.

These observations support the conclusion that the proposed measures yield much more reasonable results in assessing functional relationships between proteins within pathways, and thus these measures are more consistent with biological knowledge than existing methods.

5.4. Consistency between Gene Semantic Similarity and Protein Domain Data. Proteins are often composed of one or more functional regions, commonly referred to as protein domains [19]. Different domains typically account for different functions of proteins containing them, and thus different combinations of protein domains give rise to the diverse range of proteins found in nature. Hence, proteins can be classified into different families according to the domains that the proteins contain. Moreover, proteins containing the same domain, or say belonging to the same family, should have some similar functions and thus share some similar annotations in the biological domain of the gene ontology. Consequently, proteins belonging to the same family should have high semantic similarity scores according to the gene ontology. On the contrary, proteins belonging to different familiess should own relatively low semantic similarity scores. To assess whether the proposed similarity measures are consistent with this knowledge, we compare semantic similarity scores between proteins within a protein family and those between proteins belonging to different families as follows.

The Pfam database [20] provides a large collection of both high quality protein families (Pfam-A) and low quality protein families (Pfam-B). In version 26.0 of the Pfam-A collection (released in November 2011), 13,672 protein families are collected. From this data source, we extract 1,022 protein families, each including at least two yeast proteins. For each of these families, we calculate pairwise semantic similarity scores of proteins belonging to the family, and

we average these scores over all pairs of proteins to obtain the mean semantic similarity score within the family (ν_{in}). Meanwhile, for each protein family, we further select at random 10 times the number of proteins as those in the family, calculate semantic similarity scores between these proteins and those belonging to the family, and average over these scores to obtain the mean semantic similarity score outside the family (ν_{out}). Then, we calculate for each protein family the ratio of the mean semantic similarity scores within the family over that outside the family (ν_{in}/ν_{out}), and we average such ratios over all 1,022 protein families to obtain a criterion called fold change of semantic similarity scores within protein families against those outside families. We summarize the fold changes in Figure 4, from which we can clearly see the effectiveness of the proposed measures. For example, using the correlation measure, we obtain a fold change of 6.915. Using the cosine measure, we obtain a fold change of 6.511. Using the Jaccard measure, we obtain a fold change of 3.267. In contrast, the method of Wang only produces a fold change of 1.856, and the method of Resnik produces a slightly larger fold change of 2.370.

We further change the minimum number proteins belonging to a protein family from 2 to 10, calculate the fold change in each situation, and present the results in Table 1. Briefly, the fold change varies with the minimum number of proteins in a protein family, but the observation that the fold changes of the proposed measures are greater than those of the existing methods remains unchanged. For example, when considering protein families containing at least 10 proteins, we obtain fold changes of 9.273, 9.814, and 4.516 for the correlation, cosine, and the Jaccard measures, respectively. In contrast, the fold change for the measures of Wang, Resnik, and Schlicker are 2.090, 2.846, and 3.430, respectively. From these results, we make the conjecture that the proposed measures yield much more reasonable results in assessing functional relationships between proteins that belong to the same protein family. Hence, we conclude that the proposed measures are more consistent with biological knowledgethan existing methods.

5.5. Consistency between Gene Semantic Similarity and PPI Data. Biological knowledge suggests that proteins often interact with each other in the collaborative generation of biological functions [21]. The collection of all physical interactions in a living organism is typically referred to as the protein-protein interaction (PPI) network, in which nodes are proteins and edges are physical interactions between the proteins. Interacting proteins are usually involved in similar biological process and thus have similar annotations in the biological process domain of the gene ontology and high semantic similarity scores. To assess whether our similarity measures are consistent with this knowledge, we assess relationships between interacting proteins and their semantic similarity scores as follows.

We download two manually curated PPI networks of *Saccharomyces cerevisiae*. From BioGrid (biological generic repository for interaction datasets) [22, 23], we extract a PPI network composed of 3,529 nodes and 16,285 edges. From

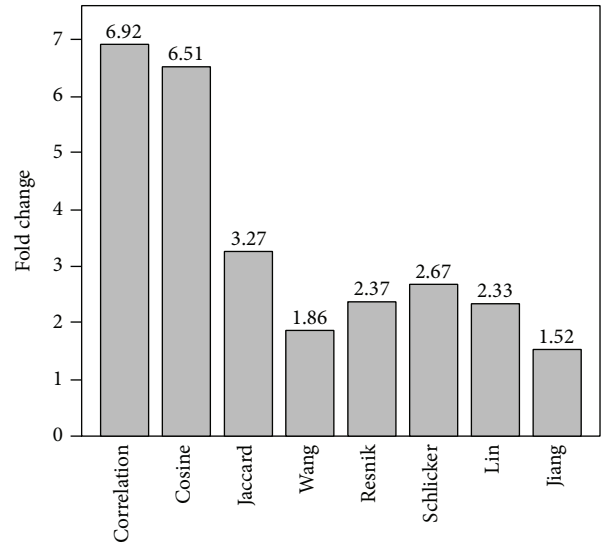


Figure 4: Fold change of semantic similarity scores within protein families against those outside protein families.

DIP (database of interacting proteins) [24, 25], we extract a relative small PPI network including 2,902 nodes and 7,005 edges. For each of these networks, we calculate semantic similarity scores for interacting proteins and those for the same number of randomly selected noninteracting pairs of proteins, and we plot the distribution of these scores in Figures 5(a) and 5(b). From the figure, we obviously see that the semantic similarity scores for interacting proteins are in general larger than those for noninteracting proteins, and this observation exists for both the BioGrid and the DIP networks.

Then, for each of these networks, we average over semantic similarity scores between interacting proteins to obtain the mean semantic similarity score of interacting proteins (τ_{int}). Meanwhile, we average over semantic similarity scores of noninteracting pairs of proteins to obtain the mean semantic similarity score of noninteracting proteins (τ_{non}). Finally, we calculate the fold change as τ_{int}/τ_{non} to measure the effectiveness of a method in distinguishing the functional relationship between interacting proteins. We present the results summarized in Figure 6, from which we can see the effectiveness of the proposed measures. For example, for the BioGrid network, we obtain a fold change of 6.15 when using the correlation measure. For the DIP network, the fold change is 5.44 for the correlation measure. For the cosine and the Jaccard measures, we observe similar results. From these observations, we make the conjecture that the semantic similarity scores calculated by the proposed measures are consistent with biological knowledge about interacting proteins.

It has also been shown that proteins closer in a PPI network tend to have more similar functions [4]. With this understanding, we use the length of the shortest path between two proteins in a PPI network to measure the network proximity of the proteins, use the semantic similarity score of the two proteins to measure their functional similarity, and

Table 1: Fold changes of semantic similarity scores within protein families against those outside families.

m	n	Semantic similarity measures							
		Correlation	Cosine	Jaccard	Wang	Resnik	Schlicker	Lin	Jiang
2	1022	6.915	6.511	3.267	1.856	2.370	2.669	2.331	1.524
3	562	8.986	8.446	3.827	1.988	2.680	3.100	2.641	1.629
4	360	9.608	8.760	4.027	2.037	2.799	3.247	2.761	1.656
5	240	9.359	9.135	4.131	2.065	2.843	3.324	2.827	1.662
6	182	9.997	9.214	4.224	2.105	2.901	3.410	2.888	1.692
7	141	10.10	9.741	4.363	2.106	2.952	3.476	2.918	1.690
8	110	9.921	9.409	4.432	2.101	2.853	3.409	2.895	1.661
9	89	9.880	9.321	4.445	2.094	2.857	3.419	2.908	1.643
10	75	9.814	9.273	4.516	2.090	2.846	3.430	2.898	1.644

m: minimum number of proteins in a family. n: number of protein families, each containing at least m proteins.

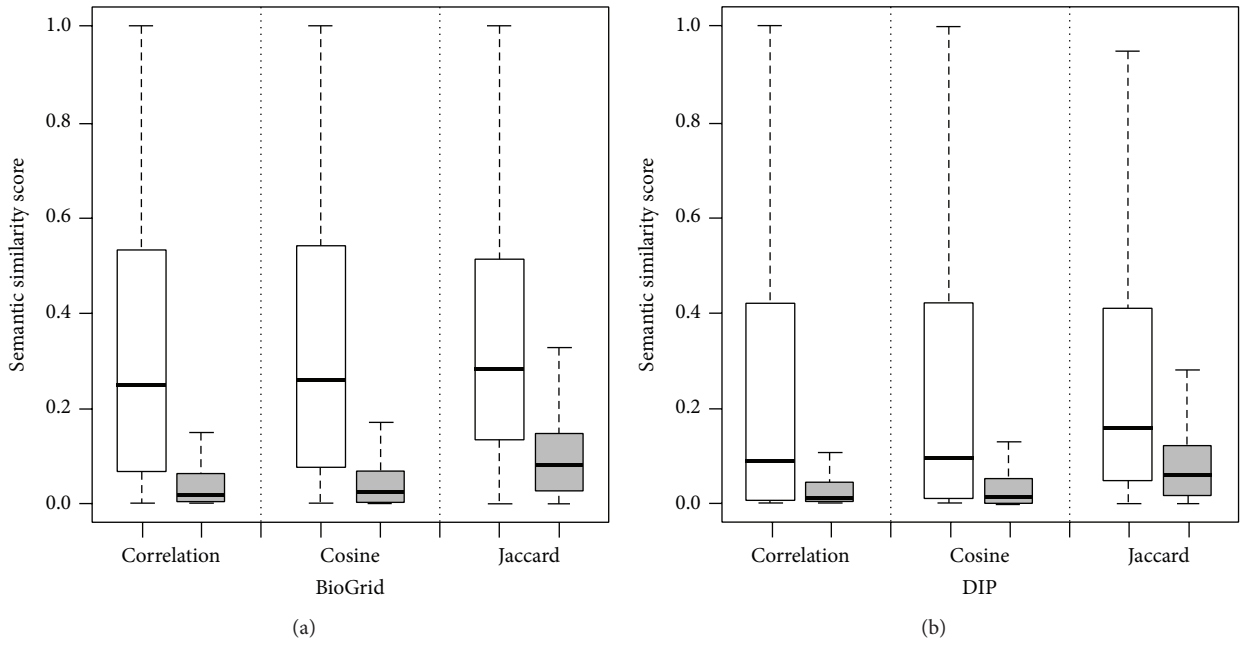


Figure 5: Relationships between semantic similarity scores and protein-protein interaction data. (a) Distributions of similarity scores of interacting proteins (white) against noninteracting proteins (gray) for the BioGrid dataset. (b) Distributions of similarity scores of interacting proteins (white) against noninteracting proteins (gray) for the DIP dataset.

plot the change of the similarity score with the closeness of proteins in Figure 6. From the figure, we can see that protein pairs tend to have higher semantic similarity scores if they are closer in the PPI network. For example, for the BioGrid network and the cosine measure, the median semantic similarity score is 0.2590 for direct interacting protein pairs, 0.0720 for protein pairs intermediated by another protein, 0.0372 for protein pairs intermediated by two other proteins, and so forth. Similar results are observed for the other two measures. These results suggest that protein similarity scores are correlated with protein closeness in a PPI network, again consistent with biological knowledge.

6. Conclusions and Discussion

In this paper, we have proposed an approach to represent annotations of a gene product in the gene ontology using vectors that are composed of information contents of terms in the ontology. Based on this notion, we have proposed to calculate pairwise semantic similarity between gene products by using three measures (Pearson's correlation coefficient, cosine similarity, and the Jaccard index) to quantify the relatedness of the corresponding vectors. We have performed a series of comprehensive studies on the effectiveness of the proposed measures using the ontology of biological process and annotations of the budding yeast

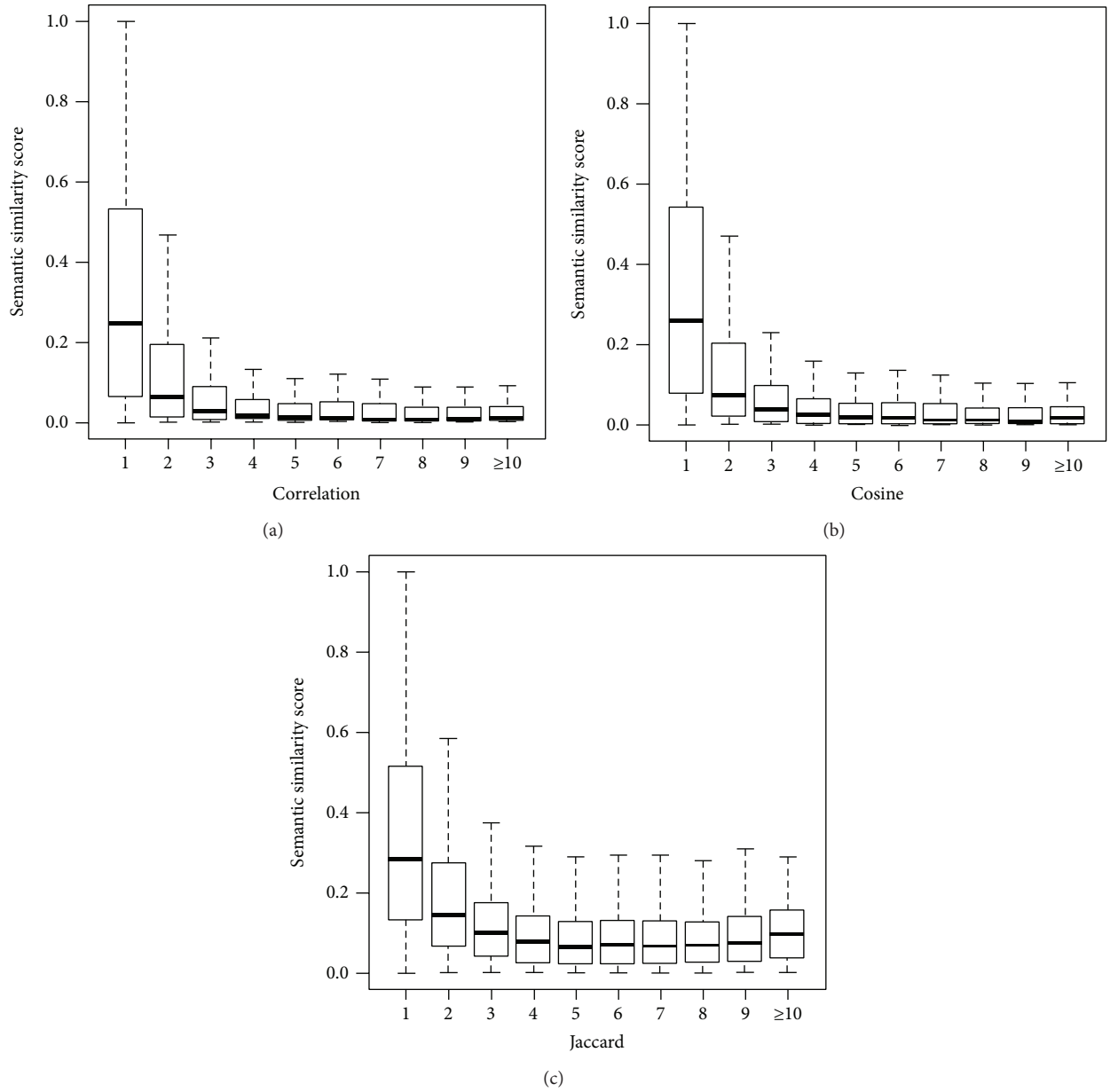


Figure 6: Distributions of semantic similarity scores against the shortest path distance of interacting proteins for the BioGrid dataset. (a) Results for the measure of correlation. (b) Results for the measure of cosine. (c) Results for the measure of Jaccard.

Saccharomyces cerevisiae. Comprehensive studies on the relationships between semantic similarity of gene products and biochemical pathways, protein families, and protein-protein interaction networks show that semantic similarity scores calculated using the proposed measures are more consistent with biological knowledge than those derived using a list of five existing methods, suggesting the effectiveness of our method in characterizing functional similarity between gene products based on the gene ontology.

The main advantage of the proposed measures is the simplicity in calculation and the effectiveness in characterizing semantic similarity between gene products. The representation of gene products as vectors of information contents of ontology terms is straightforward, making the followed computation easy to understand. The simplicity in presentation also benefits the computation with a low time complexity, thereby making our method suitable for large scale calculation of semantic similarity for not only applications based on the gene ontology but also those using other ontologies.

Certainly, the proposed measures can be further improved from the following aspects. First, although the contribution of a term in a domain ontology has been characterized by its information content, it is possible to further refine such contribution by adjusting the information contents with prior knowledge. For example, it is not hard

to combine annotations of different organisms to achieve a more precise estimation of information contents for concepts in the gene ontology. Another possibility is to develop a Bayesian method to estimate the information contents, using existing annotations to derive the prior distribution.

Second, although the presentation of domain entities as vectors of concepts is simple yet effective, the incorporation of the structure of the concepts in the underlying ontology may further improve the performance of the proposed method. Existing algorithms for calculating similarity between two tree structures [26] might be a potential candidate along this direction.

Conflict of Interests

The author does not have any conflict of interests.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (no. 71101010).

References

[1] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[2] R. Jiang, M. Gan, and P. He, "Constructing a gene semantic similarity network for the inference of disease genes," *BMC Systems Biology*, vol. 5, supplement 2, article S2, 2011.

[3] J. Wu, Y. Li, and R. Jiang, "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies," *PLoS Genetics*, vol. 10, no. 3, Article ID e1004237, 2014.

[4] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, "Linking human diseases to animal models using ontology-based phenotype annotation," *PLoS Biology*, vol. 7, no. 11, Article ID e1000247, 2009.

[5] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[6] M. Courtot, N. Juty, C. Knupfer et al., "Controlled vocabularies and semantics in systems biology," *Molecular Systems Biology*, vol. 7, p. 543, 2011.

[7] M. Gan, X. Dou, and R. Jiang, "From ontology to semantic similarity: calculation of ontology-based semantic similarity," *The Scientific World Journal*, vol. 2013, Article ID 793091, 11 pages, 2013.

[8] Y. Chen, J. Hao, W. Jiang et al., "Identifying potential cancer driver genes by genomic data integration," *Scientific Reports*, vol. 3, article 3538, 2013.

[9] Y. Chen, X. Wu, and R. Jiang, "Integrating human omics data to prioritize candidate genes," *BMC Medical Genomics*, vol. 6, no. 1, article 57, 2013.

[10] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[11] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.

[12] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304, Morgan Kaufmann, 1998.

[13] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics*, pp. 19–33, 1997.

[14] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC Bioinformatics*, vol. 7, article 302, 2006.

[15] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, Article ID btq064, pp. 976–978, 2010.

[16] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene Ontology terms," *Data and Knowledge Engineering*, vol. 61, no. 1, pp. 137–152, 2007.

[17] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[18] J. M. Cherry, E. L. Hong, C. Amundsen et al., "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, pp. D700–D705, 2012.

[19] A. Bateman, L. Coin, R. Durbin et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 32, pp. D138–D141, 2004.

[20] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[21] R. Jiang, Z. Tu, T. Chen, and F. Sun, "Network motif identification in stochastic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 25, pp. 9404–9409, 2006.

[22] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, no. 1, pp. D698–D704, 2011.

[23] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, pp. D535–D539, 2006.

[24] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

[25] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2000.

[26] Y. Zhong, C. A. Meacham, and S. Pramanik, "A general method for tree-comparison based on subtree similarity and its use in a taxonomic database," *Biosystems*, vol. 42, no. 1, pp. 1–8, 1997.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 761562, 7 pages
http://dx.doi.org/10.1155/2014/761562

Research Article

State Observer Design for Delayed Genetic Regulatory Networks

Li-Ping Tian,[1] Zhi-Jun Wang,[2] Amin Mohammadbagheri,[3] and Fang-Xiang Wu[3,4]

[1] *School of Information, Beijing Wuzi University, Beijing 101149, China*
[2] *College of Mathematics and Statistics, Hebei University of Economics and Business, Shijiazhuang, Hebei 050061, China*
[3] *Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*
[4] *Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*

Correspondence should be addressed to Fang-Xiang Wu; faw341@mail.usask.ca

Received 14 March 2014; Accepted 6 May 2014; Published 22 May 2014

Academic Editor: Zhongming Zhao

Genetic regulatory networks are dynamic systems which describe the interactions among gene products (mRNAs and proteins). The internal states of a genetic regulatory network consist of the concentrations of mRNA and proteins involved in it, which are very helpful in understanding its dynamic behaviors. However, because of some limitations such as experiment techniques, not all internal states of genetic regulatory network can be effectively measured. Therefore it becomes an important issue to estimate the unmeasured states via the available measurements. In this study, we design a state observer to estimate the states of genetic regulatory networks with time delays from available measurements. Furthermore, based on linear matrix inequality (LMI) approach, a criterion is established to guarantee that the dynamic of estimation error is globally asymptotically stable. A gene repressillatory network is employed to illustrate the effectiveness of our design approach.

1. Introduction

Recently nonlinear differential equations have been proposed to model genetic regulatory networks. Based on this model, stability of genetic regulatory networks has been intensively studied, which is believed useful in designing and controlling genetic regulatory networks. In [1], sufficient and necessary local delay-independent stability conditions are given for several types of simplified genetic regulatory networks with a single time delay. In [2, 3], we present some sufficient and necessary conditions of local delay-independent stability conditions for general genetic regulatory networks with a single time delay and multiple time delays. Some sufficient conditions for global stability of genetic regulatory networks have been derived based on LMI approaches [4–6] and M-matrix theorem [7, 8].

On the other hand, to understand the dynamic behavior of genetic regulatory networks, measurements of all internal states are very useful. The internal states of a genetic regulatory network consist of the concentrations of mRNA and proteins involved in it. However, because of some limitations such as experiment techniques, not all internal states of genetic regulatory network can be effectively measured. As a result, the internal states of genetic regulatory networks cannot be completely available. Therefore, the state estimation problem can play an important role in understanding the dynamic behaviors of genetic regulatory networks. The state estimation problem addressed is to estimate the states based on available output measurements such that the dynamic of estimation error is globally asymptotically stable. Actually, the state estimation methods have been very important in understanding, designing, and controlling dynamic systems such as engineering control system [9], neural networks [10, 11], and complex systems [12].

In this study, we will study the state estimation of genetic regulatory networks with time delays modeled by nonlinear differential equations. Section 2 briefly describes delayed genetic regulatory networks with SUM regulatory logic. In Section 3 we design a full-order state observer to estimate the states of delayed genetic regulatory networks. Some properties of this observer are discussed. In Section 4, based on LMI approach we establish a sufficient condition under which the dynamic of estimation error for designed state observer is asymptotically and delay-independently stable.

In Section 5, a gene repressillatory network is employed to illustrate the effectiveness of our approach described in Section 4. Section 6 gives our conclusion of this study and points out some directions of future work.

2. Delayed Genetic Regulatory Networks

A delayed genetic regulatory network consisting of n mRNAs and n proteins can be described by the following equations:

$$\begin{aligned} \dot{m}_i(t) &= -k_{mi} m_i(t) + c_i\left(p\left(t-\tau_p\right)\right) \\ \dot{p}_i(t) &= -k_{pi} p_i(t) + r_i m_i\left(t-\tau_m\right) \\ &\qquad \text{for } i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where $m_i(t)$, $p_i(t) \in R_+^n$ represent the concentrations of mRNA i and protein i, respectively. k_{mi} and k_{pi} are positive real numbers that represent the degradation rates of mRNA i and protein i, respectively. r_i is a positive constant representing the rate of translating mRNA i to protein i. $c_i(p(t, \tau_p))$ is a nonlinear function of $p_1(t-\tau_p), \ldots, p_n(t-\tau_p)$ representing the regulation function of gene i. Both τ_m and τ_p are positive constants indicating time delays of mRNAs and proteins, respectively.

The bottom equation in model (1) describes the translational process. The term $r_i m_i(t)$ reflects the fact that one kind of proteins is translated only from one kind of mRNA molecules. The top equation in model (1) describes the transcriptional process. One gene or mRNA is generally activated or repressed by multiple proteins in the transcriptional process indicated in the definition of $c_i(p(t))$. In this paper, we take $c_i(p(t)) = \sum_{j=1}^{n} c_{ij}(p_j(t))$, which is called the "SUM" logic [13]. That is, each transcription factor acts additively to regulate gene i. The SUM logic is applicable if one gene can be regulated by several proteins independently by binding with different promoters or by a family of similar proteins independently binding to one promoter. In many natural gene networks, this SUM logic does exist [13]. The regulation function $c_{ij}(p_j(t))$ is a function of the Hill form [14] as follows:

$$c_{ij}\left(p_j(t)\right) = a_{ij} \frac{1}{1 + \left(p_j(t)/b_j\right)^{h_j}} \tag{2}$$

if transcription factor j is a repressor of gene i, or

$$c_{ij}\left(p_j(t)\right) = a_{ij} \frac{\left(p_j(t)/b_j\right)^{h_j}}{1 + \left(p_j(t)/b_j\right)^{h_j}} \tag{3}$$

if transcription factor j is an activator of gene i, where a_{ij} and b_j are nonnegative constants and h_j is the Hill coefficient representing the degree of cooperativity. In this study, assume that $h_j \geq 1$. Note that

$$\frac{1}{1 + \left(p_j(t)/b_j\right)^{h_j}} = 1 - \frac{\left(p_j(t)/b_j\right)^{h_j}}{1 + \left(p_j(t)/b_j\right)^{h_j}}. \tag{4}$$

Then system (1) can be rewritten as follows:

$$\begin{aligned} \dot{m}(t) &= -K_m m(t) + G g\left(p\left(t-\tau_p\right)\right) + L \\ \dot{p}(t) &= -K_p p(t) + R m\left(t-\tau_m\right), \end{aligned} \tag{5}$$

where $m(t) = (m_1(t), \ldots, m_n(t))$ and $p(t) = (p_1(t), \ldots, p_n(t))$; $K_m = \mathrm{diag}(k_{m_1}, \ldots, k_{m_n})$, $K_p = \mathrm{diag}(k_{p_1}, \ldots, k_{p_n})$, and $R = \mathrm{diag}(r_1, \ldots, r_n)$; $G = (G_{ij})$ is an $n \times n$ stoichiometric matrix representing regulatory relationships of the network, which is defined as follows: $G_{ij} = 0$ if transcription factor j does not directly regulate gene i, $G_{ij} = a_{ij}$ if transcription factor j directly activates gene i, and $G_{ij} = -a_{ij}$ if transcription factor j directly represses gene i; $L = (l_1, \ldots, l_n)$ where l_i is a constant and is defined as $l_i = \sum_{j \in \mathrm{Rep}} a_{ij}$, where Rep is the set of repressors of gene i. $g = (g_1, \ldots, g_n)$ where $g_j(u) = (u/b_j)^{h_j}/[1 + (u/b_j)^{h_j}]$ is a monotonically increasing function. Obviously these functions with $h_j \geq 1$ have the continuous derivatives for $u \geq 0$. From calculus, we have

$$\theta_j = \max_{u \geq 0} g_j'(u) = \frac{\left(h_j - 1\right)^{(h_j-1)/h_j}\left(h_j + 1\right)^{(h_j+1)/h_j}}{4 b_j h_j} > 0. \tag{6}$$

3. State Observer

In practice, the information about the network states is often incomplete from the experimental measurements. For example, the concentrations of proteins might be immeasurable because of the limitation of measurement techniques. Our purpose of this study is to develop an efficient estimation system (called a state observer) in order to estimate the network states from the available measurements. In this paper, assume that measurements are the linear combinations of mRNA and protein concentrations and thus the output can be expressed as follows:

$$z(t) = C \begin{bmatrix} m(t) \\ p(t) \end{bmatrix}, \tag{7}$$

where $z(t)$ is an m-dimensional vector representing the measurements and C is an $m \times 2n$ observation matrix. Unless the rank of matrix C in (7) is $2n$, the states of system (5) cannot be exactly estimated from the static observation equation (7) only. In practice, the rank of matrix C in (7) is less than $2n$. To approximately estimate the states of a dynamic system, a dynamic system similar to the original one is designed to estimate the states. In this paper, the full-order state estimator of network (5) is designed as follows:

$$\begin{aligned} \begin{bmatrix} \dot{\widehat{m}}(t) \\ \dot{\widehat{p}}(t) \end{bmatrix} &= -\begin{bmatrix} K_m & 0 \\ 0 & K_p \end{bmatrix} \begin{bmatrix} \widehat{m}(t) \\ \widehat{p}(t) \end{bmatrix} + \begin{bmatrix} G g\left(\widehat{p}\left(t-\tau_p\right)\right) + L \\ R \widehat{m}\left(t-\tau_m\right) \end{bmatrix} \\ &\quad + D\left(z(t) - C \begin{bmatrix} \widehat{m}(t) \\ \widehat{p}(t) \end{bmatrix}\right), \end{aligned} \tag{8}$$

where $\widehat{m}(t)$ and $\widehat{p}(t)$ are the estimation of states and D is $2n \times m$ estimate gain matrix to be determined.

Let the estimation error be

$$x(t) = m(t) - \widehat{m}(t) \qquad y(t) = p(t) - \widehat{p}(t). \tag{9}$$

Then from (5), (8), and (9), the error system can be described as follows:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} = -(K + DC) \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} + \begin{bmatrix} Gf\left(t - \tau_p\right) \\ Rx\left(t - \tau_m\right) \end{bmatrix}, \tag{10}$$

where $K = \mathrm{diag}(K_m, K_p)$ and $f(t) = g(p(t)) - g(\widehat{p}(t))$.

Now designing the state estimator for network (5) is reduced to find the estimate gain matrix D such that the error system (10) is globally asymptotically stable. From (6), we have

$$0 \le \frac{f_j(t)}{y_j(t)} \le \theta_j \qquad \text{for } j = 1, 2, \ldots, n. \tag{11}$$

Furthermore, for any nonnegative diagonal $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \ge 0$, from (11) it follows that

$$\begin{aligned} &-f^T\left(t - \tau_p\right) 2\Lambda f\left(t - \tau_p\right) + y^T\left(t - \tau_p\right) 2\Lambda\Theta f\left(t - \tau_p\right) \\ &\quad \ge 0, \end{aligned} \tag{12}$$

where $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$.

Once matrix D is determined, the estimations $\widehat{m}(t)$ and $\widehat{p}(t)$ are numerically calculated from (8). That the same technique can be applied for solving (1) directly is the same as solving system (8) with $D = 0$, which results in an estimation error system (10) with $D = 0$. If the system (1) is unstable and the values of $\widehat{m}(0)$ and $\widehat{p}(0)$ are different from their true counterparts, then the estimation errors will be exponentially increased. Even if the values of $\widehat{m}(0)$ and $\widehat{p}(0)$ are the exact same as their true counterparts, the round-off errors can also cause the estimation errors to be exponentially increased. Therefore, in practice it is important to design matrix D to make sure the estimation error system is stable. Theorems 1 and 2 in next section will guarantee that, for any values of $\widehat{m}(0)$ and $\widehat{p}(0)$, the estimation errors will be asymptotically converged to zero.

4. Main Results and Proofs

In this section we will first derive the conditions under which the error system (10) is globally asymptotically stable for a given estimate gain matrix.

Theorem 1. *For a given estimate gain matrix D, the error system* (10) *has a unique equilibrium state $x = 0$ and $y = 0$ and is globally asymptotically stable if there exist $2n \times 2n$ positive definite matrices P and $n \times n$ positive definite matrices Q and S and positive diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) > 0$, such that the following LMI holds:*

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & 0 \\ \Omega_{12}^T & \Omega_{22} & \Omega_{23} \\ 0 & \Omega_{23}^T & -S \end{bmatrix} < 0, \tag{13}$$

where $\Omega_{11} = -(K + DC)^T P - P(K + DC) + \mathrm{diag}(Q, S)$, $\Omega_{22} = -\mathrm{diag}(2\Lambda, Q)$, $\Omega_{12} = P\,\mathrm{diag}(G, R)$, and $\Omega_{23} = [\Lambda\Theta, 0]^T$.

Proof. Consider the following Lyapunov-Krasovskii functional:

$$V\left(x(t), y(t)\right) = V_1\left(x(t), y(t)\right) + V_2\left(x(t), y(t)\right), \tag{14}$$

where

$$\begin{aligned} V_1\left(x(t), y(t)\right) &= \begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} P \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \\ V_2\left(x(t), y(t)\right) &= \int_{t-\tau_m}^{t} x^T(u) Q x(u)\, du \\ &\quad + \int_{t-\tau_p}^{t} y^T(u) S y(u)\, du. \end{aligned} \tag{15}$$

Differentiating $V_i(x(t), y(t))$ defined above along the trajectories of system (10), we have

$$\begin{aligned} \dot{V}_1\left(x(t), y(t)\right) &= 2\begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} P \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} \\ &= -2\begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} P(K + DC) \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ &\quad + 2\begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} \\ &\quad \times P\,\mathrm{diag}(G, R) \begin{bmatrix} f\left(t - \tau_p\right) \\ x\left(t - \tau_m\right) \end{bmatrix}, \\ \dot{V}_2\left(x(t), y(t)\right) &= x^T(t) Q x(t) - x^T\left(t - \tau_m\right) Q x\left(t - \tau_m\right) \\ &\quad + y^T(t) S y(t) - y^T\left(t - \tau_p\right) S y\left(t - \tau_p\right). \end{aligned} \tag{16}$$

Taking inequality (12) into consideration, we have

$$\begin{aligned} &\dot{V}\left(x(t), y(t)\right) \\ &\le \dot{V}_1\left(x(t), y(t)\right) + \dot{V}_2\left(x(t), y(t)\right) \\ &\quad - f^T\left(t - \tau_p\right) 2\Lambda f\left(t - \tau_p\right) \\ &\quad + y^T\left(t - \tau_p\right) 2\Lambda\Theta f\left(t - \tau_p\right) \\ &= \begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} \left(-2P(K + DC) + \mathrm{diag}(Q, S)\right) \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ &\quad + 2\begin{bmatrix} x^T(t) & y^T(t) \end{bmatrix} P\,\mathrm{diag}(G, R) \begin{bmatrix} f\left(t - \tau_p\right) \\ x\left(t - \tau_m\right) \end{bmatrix} \\ &\quad - \begin{bmatrix} f^T\left(t - \tau_p\right) & x^T\left(t - \tau_m\right) \end{bmatrix} \mathrm{diag}(2\Lambda, Q) \begin{bmatrix} f\left(t - \tau_p\right) \\ x\left(t - \tau_m\right) \end{bmatrix} \end{aligned}$$

$$+ 2y^T\left(t-\tau_p\right)\left[\Lambda\Theta \;\; 0\right]\begin{bmatrix} f\left(t-\tau_p\right) \\ x\left(t-\tau_m\right)\end{bmatrix}$$
$$- y^T\left(t-\tau_p\right)Sy\left(t-\tau_p\right)$$
$$= \xi^T(t)\,\Omega\xi(t) < 0, \tag{17}$$

where $\xi(t) = [(x^T(t), y^T(t)), (f^T(t-\tau_p), x^T(t-\tau_p)), y^T(t-\tau_p)]^T$.

From Lyapunov-Krasovskii theory [15], the error system (10) is globally asymptotically stable. From (10), $x = 0$ and $y = 0$ are an equilibrium state. To prove the uniqueness of the equilibrium state of the error system (10), here we use proof-by-contradiction technique. Note that Lyapunov-Krasovskii functional (18) associated with the error system (10) is independent of the equilibrium state. Therefore if the error system (10) has another equilibrium state, it is also globally asymptotically stable, which is not possible.

In Theorem 1, for a given estimate gain matrix D, the stability condition of the error dynamic system (10) is established in terms of linear matrix inequality (LMI) which can be solved by standard MATLAB function. If matrix D is unknown, matrix inequality (13) becomes nonlinear in matrices, P, D, Q, S, and Λ, which is not easy to be solved. However, let $PD = -T$; then matrix inequality (13) becomes linear in matrices, P, T, Q, S, and Λ. Therefore, we have the following theorem. □

Theorem 2. *If there exist $2n \times 2n$ positive definite matrices P and $n \times n$ positive definite matrices Q and S, positive diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) > 0$, and an $2n \times m$ matrix T such that the following LMI*

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & 0 \\ \Omega_{12}^T & \Omega_{22} & \Omega_{23} \\ 0 & \Omega_{23}^T & -S \end{bmatrix} < 0 \tag{18}$$

holds, where $\Omega_{11} = -KP - PK + C^TT^T + TC + \mathrm{diag}(Q, S)$, and sub-matrices Ω_{22}, Ω_{12}, and Ω_{23} are the same as in Theorem 1, then with the estimator gain matrix

$$D = -P^{-1}T \tag{19}$$

the error system (10) *has a unique equilibrium state $x = 0$ and $y = 0$ and is globally asymptotically stable.*

Proof of Theorem 2 is straightforward from Theorem 1 and thus is omitted here.

5. An Illustration Example

In this section, we employ the gene repressilatory network to show the effectiveness and correctness of our theoretical results. The gene repressilatory network consists of three genes and three proteins (*lacl*, *tetR*, and *cl*), each repressing the transcription of its downstream partner [16] as shown in Figure 1. This network without time delays has been studied theoretically and experimentally in [16]. The delay-independent local and global stability of this gene repressilatory network with time delays has widely been studied in [1–8].

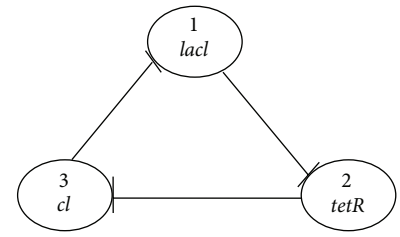


Figure 1: Structure of gene repressilatory network.

The mathematical model of this gene repressilatory network with time delay is described by the following equation:

$$\dot{m}_i(t) = -k_m m_i(t) + \frac{a}{1 + p_{i-1}^h\left(t-\tau_p\right)},$$
$$\dot{p}_i(t) = -k_p p_i(t) + r m_i\left(t-\tau_m\right), \tag{20}$$

where k_m, a, k_p, and r are positive constants and subscript $0 = 3$.

In this study we consider gene repressillatory network (20) with the values of parameters set as follows: $h = 2$, $k_m = 1.2$, $a = 2.5$, $k_p = 1$, and $r = 0.8$. For system (20) with these parameter specifications, we have

$$K_m = \begin{bmatrix} 1.2 & 0 & 0 \\ 0 & 1.2 & 0 \\ 0 & 0 & 1.2 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & -2.5 & 0 \\ 0 & 0 & -2.5 \\ -2.5 & 0 & 0 \end{bmatrix},$$
$$K_p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}. \tag{21}$$

And $\theta_j = 3\sqrt{3}/8$ for $j = 1, 2, 3$.

Case A. Assume that the concentration of all proteins is unable to be measured. The observation matrix C is

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{22}$$

By using MATLAB LMI toolbox, we solve LMIs (18) with the above data for P, T, Q, S, and Λ and obtain

$$P = \begin{bmatrix} 1.8750 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.8750 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.8750 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8.3807 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8.3807 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8.3807 \end{bmatrix},$$

$$T = -\begin{bmatrix} 10.1248 & 0 & 0 \\ 0 & 10.1248 & 0 \\ 0 & 0 & 10.1248 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} 9.2810 & 0 & 0 \\ 0 & 9.2810 & 0 \\ 0 & 0 & 9.2810 \end{bmatrix},$$

$$S = \begin{bmatrix} 7.9497 & 0 & 0 \\ 0 & 7.9497 & 0 \\ 0 & 0 & 7.9497 \end{bmatrix},$$

$$\Lambda = \begin{bmatrix} 6.9963 & 0 & 0 \\ 0 & 6.9963 & 0 \\ 0 & 0 & 6.9963 \end{bmatrix}. \tag{23}$$

Therefore, we have

$$D = -P^{-1}T = \begin{bmatrix} 5.4 & 0 & 0 \\ 0 & 5.4 & 0 \\ 0 & 0 & 5.4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{24}$$

Figure 2 depicts the estimation errors of protein concentrations of delayed genetic regulatory network (20) with specified parameters in the caption of Figure 2. From Figure 2, it can be seen that in six minutes the estimated protein concentrations are exactly the same as the true protein concentrations although they are not measured. The time that needs to exactly estimate the true states depends on the initial errors between the true states and estimated states (which are random guesses in practice). In Figure 2, the initial errors of protein estimations range from 0.1 to 0.6. If the initial errors are zero, the estimated state would be the exact true states from beginning on.

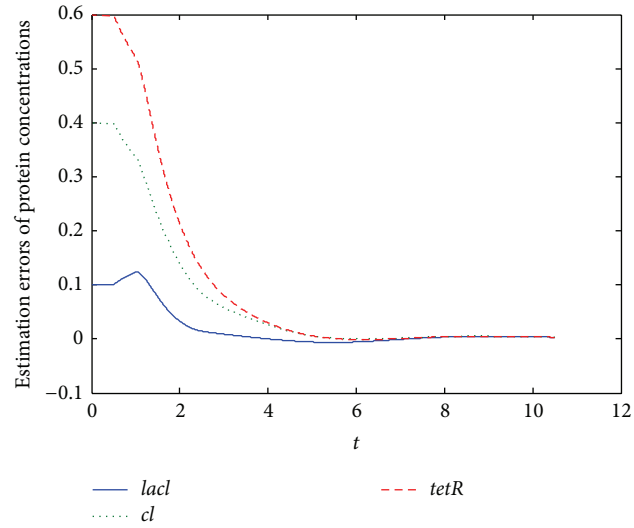


Figure 2: Estimation errors of protein concentrations of system (20) with specified parameters and $\tau_p = \tau_m = 0.5$ minutes while mRNA concentrations are available.

Case B. Assume that the concentration of all proteins is able to be measured while we would like to estimate the gene expressions. The observation matrix C becomes

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{25}$$

By using MATLAB LMI toolbox, we solve LMIs (18) with above data for P, T, Q, S, and Λ and obtain

$$P = \begin{bmatrix} 19.3747 & 0 & 0 & 0 & 0 & 0 \\ 0 & 19.3747 & 0 & 0 & 0 & 0 \\ 0 & 0 & 19.3747 & 0 & 0 & 0 \\ 0 & 0 & 0 & 21.9879 & 0 & 0 \\ 0 & 0 & 0 & 0 & 21.9879 & 0 \\ 0 & 0 & 0 & 0 & 0 & 21.9879 \end{bmatrix},$$

$$T = -\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 53.1451 & 0 & 0 \\ 0 & 53.1451 & 0 \\ 0 & 0 & 53.1451 \end{bmatrix},$$

$$Q = \begin{bmatrix} 17.6592 & 0 & 0 \\ 0 & 17.6592 & 0 \\ 0 & 0 & 17.6592 \end{bmatrix},$$

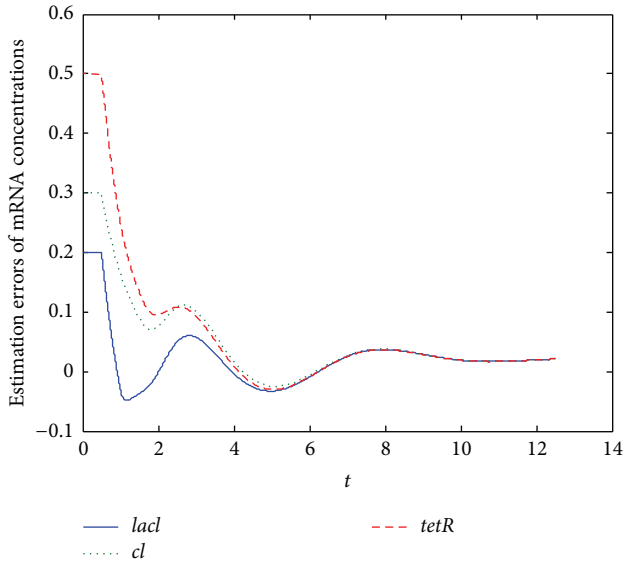


FIGURE 3: Estimation errors of mRNA concentrations of system (20) with specified parameters and $\tau_p = \tau_m = 0.5$ minutes while protein concentrations are available.

$$S = \begin{bmatrix} 75.9017 & 0 & 0 \\ 0 & 75.9017 & 0 \\ 0 & 0 & 75.9017 \end{bmatrix},$$
$$\Lambda = \begin{bmatrix} 69.2189 & 0 & 0 \\ 0 & 69.2189 & 0 \\ 0 & 0 & 69.2189 \end{bmatrix}. \tag{26}$$

Therefore, we have

$$D = -P^{-1}T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2.4170 & 0 & 0 \\ 0 & 2.4170 & 0 \\ 0 & 0 & 2.4170 \end{bmatrix}. \tag{27}$$

Figure 3 depicts the estimation errors of mRNA concentrations of delayed genetic regulatory network (20) with specified parameters in the caption of Figure 3 and knowing protein concentrations. From Figure 3, it can be seen that in about ten minutes the estimated mRNA concentrations can pretty well approximate the true mRNA concentrations although they are not measured. In Figure 3, the initial errors of protein estimations range from 0.2 to 0.5.

6. Conclusion and Future Work

In this paper, we have studied the state estimation of genetic regulatory networks with time delays. Based on LMI approach, a full-order state observer is designed to estimate the states from incomplete measurements so that the state estimation error is globally asymptotically stable. The theorems presented in this paper have been illustrated by the gene repressillatory network. The simulation results have verified that our designed observer can effectively estimate the unmeasured states. In this study, we assume that all parameters of genetic regulatory networks are available. In practice, some of parameters in networks may be unknown. One direction of our future work is to employ the extended Kalman filter [17] to estimate the known parameters and state of the systems simultaneously. Parameter uncertainties and noise perturbations exist in genetic regulatory networks [4, 6, 7, 16, 18] and measured outputs, which can affect the performance of state observer. The second direction of our future work is to design robust state observer for genetic regulatory networks with parameter uncertainties and noises. Typically measured outputs are sampled at a series of time points although state variables of genetic regulatory networks are continuous. The third direction of our future work is to design a state observer for genetic regulatory networks with discretized outputs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Base Fund of Beijing Wuzi University and Fund for Beijing Excellent Team for Teaching Mathematics through LPT and by Natural Sciences and Engineering Research Council of Canada (NSERC) through FXW. The authors would like to thank the reviewers for their comments and suggestions.

References

[1] L. Chen and K. Aihara, "Stability of genetic regulatory networks with time delay," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no. 5, pp. 602–608, 2002.

[2] F.-X. Wu, "Delay-independent stability of genetic regulatory networks with time delays," *Advances in Complex Systems*, vol. 12, no. 1, pp. 3–19, 2009.

[3] F.-X. Wu, "Stability analysis of genetic regulatory networks with multiple time delays," in *Proceedings of the 29th Annual International Conference of IEEE-EMBS, Engineering in Medicine and Biology Society (EMBC '07)*, pp. 1387–1390, fra, August 2007.

[4] C. Li, L. Chen, and K. Aihara, "Stability of genetic networks with SUM regulatory logic: Lur'e system and LMI approach," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2451–2458, 2006.

[5] C. Li, L. Chen, and K. Aihara, "Synchronization of coupled nonidentical genetic oscillators," *Physical Biology*, vol. 3, no. 1, pp. 37–44, 2006.

[6] F. Ren and J. Cao, "Asymptotic and robust stability of genetic regulatory networks with time-varying delays," *Neurocomputing*, vol. 71, no. 4–6, pp. 834–842, 2008.

[7] F.-X. Wu, "Global and robust stability analysis of genetic regulatory networks with time-varying delays and parameter uncertainties," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 4, pp. 391–398, 2011.

[8] F.-X. Wu, "Delay-independent stability of genetic regulatory networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1685–1693, 2011.

[9] A. Sinha, *Linear Systems: Optimal and Robust Control*, CRC Press, New York, NY, USA, 2007.

[10] Z. Wang, D. W. C. Ho, and X. Liu, "State estimation for delayed neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 279–284, 2005.

[11] Y. He, Q.-G. Wang, M. Wu, and C. Lin, "Delay-dependent state estimation for delayed neural networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 1077–1081, 2006.

[12] Y. Y. Liu, J. J. Slotine, and A. L. Barabasi, "Observability of complex systems," *Proceedings of the National Academy of Sciences of the USA*, vol. 110, no. 7, pp. 2460–2465, 2013.

[13] S. Kalir, S. Mangan, and U. Alon, "A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*," *Molecular Systems Biology*, vol. 1, 2005.

[14] J. Nielsen, J. Villadsen, and G. Liden, *Bioreaction Engineering Principles*, Kluwer Academic/Plenum Publishers, New York, NY, USA, 2nd edition, 2003.

[15] V. B. Kolmanovskii and A. D. Myshkis, *Introduction to the Theory and Applications of Functional Differential Equations*, Kluwer Academic, Dodrecht, The Netherlands, 1999.

[16] M. B. Elowitz and S. Leibier, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.

[17] D. Simon, *Optimal State Estimation: Kalman, H_∞ and Nonlinear Approaches*, Wiley-Interscience, New Jersey, NJ, USA, 2006.

[18] L. P. Tian, Z. K. Shi, L. Z. Liu, and F. X. Wu, "M-Matrix based stability conditions for genetic regulatory networks with time-varying delays and noise perturbations," *IET Systems Biology*, vol. 7, no. 5, pp. 214–222, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 203871, 8 pages
http://dx.doi.org/10.1155/2014/203871

Research Article

DV-Curve Representation of Protein Sequences and Its Application

Wei Deng[1,2] and Yihui Luan[1]

[1] *School of Mathematics, Shandong University, Jinan 250100, China*
[2] *School of Science, Shandong Jianzhu University, Jinan 250101, China*

Correspondence should be addressed to Yihui Luan; yhluan@sdu.edu.cn

Received 7 January 2014; Revised 10 March 2014; Accepted 3 April 2014; Published 8 May 2014

Academic Editor: Rui Jiang

Based on the detailed hydrophobic-hydrophilic(HP) model of amino acids, we propose dual-vector curve (DV-curve) representation of protein sequences, which uses two vectors to represent one alphabet of protein sequences. This graphical representation not only avoids degeneracy, but also has good visualization no matter how long these sequences are, and can reflect the length of protein sequence. Then we transform the 2D-graphical representation into a numerical characterization that can facilitate quantitative comparison of protein sequences. The utility of this approach is illustrated by two examples: one is similarity/dissimilarity comparison among different ND6 protein sequences based on their DV-curve figures the other is the phylogenetic analysis among coronaviruses based on their spike proteins.

1. Introduction

The graphical representation method has become very common to analyze the huge amount of gene data. Generally, with this method we can first observe visual qualitative inspection in order to recognize major differences among similar gene sequences and further draw some mathematical characterizations of sequences to analyze their similarity/dissimilarity and evolutionary homology.

Letter sequence representation (LSR) of DNA sequences represents each base by a letter of four different letters such as A, T, G, and C. DNA sequences can be represented in different dimension spaces. For example, G-curve and H-curve [1] were first proposed by Hamori and Ruskin before thirty years. Later, Gates [2] established a 2D graphical representation that was simpler than H curve. However, Gate's graphical representation has high degeneracy because of some circuits appearing in its curve. Several researchers in their recent studies have outlined different kinds of DNA sequences graphical representation based on 2D [3–11], 3D [12–15], 4D [16], 5D [17], and 6D [18]. Among these methods, we here stress DV-curve representation which was proposed by Zhang [10]. DV-curve uses two vectors to represent one alphabet of DNA sequences and avoids degeneracy and loss of information. Furthermore, DV-curve has good visualization no matter how long these sequences are and can reflect the length of the DNA sequence.

LSR of protein sequences represents each amino acid by a letter of twenty different letters such as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V. Although protein sequences and DNA sequences belong to symbolic sequences, the methods for the graphical representation of protein sequences are relatively less popular, compared with DNA sequences. The key reason is that the extension of DNA graphical representation to protein sequences enormously increases the number of possible alternative assignments for these 20 amino acids. The amino acid sequence is the key to discover protein structure and function in the cell, so analysis of amino acid sequences is a very important part of postgenomic studies. The graphical representation study of protein sequences emerged very recently. The first visualization protein model was proposed by Randić et al. until 2004 [19]. Some researchers have studied on graphical representation of protein sequences from different perspectives [20–29].

In this paper, we introduce DV-curve graphical representation of protein sequences based on the detailed hydrophobic-hydrophilic (HP) model of amino acids. According to the important hydropathy, this approach is accompanied by a relatively small number of arbitrary choices associated with the graphical representation of proteins. Also, this representation has relatively good visualization effect to describe protein sequences in a perceivable way. As its application, we analyze the similarity/dissimilarity among some ND6 sequences and construct the phylogenetic tree of 35 coronavirus spike proteins.

2. DV-Curve Representation of Protein Sequences

2.1. Classification of Protein Sequences. The amino acid sequence is closely related to biological function. The closer the genetic relationship is, the smaller the difference in amino acid composition between them will be. Over the past thirty years, the characteristics of protein sequences have been studied by establishing different classified models [21–24, 26, 27]. A well-known model of protein sequences is the hydrophobic (H or nonpolar)-hydrophilic (P or polar), that is, the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues [30]. Based on Brown's work [31], 20 different kinds of amino acids are divided into four groups: nonpolar (np), negative polar (nep), uncharged polar (up), and positive polar (pp). This is called the detailed HP model, which can provide more information than the original HP model.

For a given protein sequence $S = S_1S_2\cdots S_n$ with length n, where S_i is the letter in the ith position among the protein sequence ($i = 1, 2, \ldots, n$), we define a primary protein sequence as a symbolic sequence which includes four letters according to the following rule:

$$b_i = \begin{cases} B_1, & \text{if } S_i \in \text{np}, \\ B_2, & \text{if } S_i \in \text{nep}, \\ B_3, & \text{if } S_i \in \text{up}, \\ B_4, & \text{if } S_i \in \text{pp}. \end{cases} \tag{1}$$

So b_i is the substitution for S_i, and then we obtain a sequence $G(s) = b_1b_2\cdots b_n$. Here b_i is a letter of the alphabet B_1, B_2, B_3, B_4. For example, for a given protein primary sequence $S = WTFESRNDPAK$, we can transform it into a new sequence according to the above rule, $G(S) = B_1B_3B_1B_2B_3B_4B_3B_2B_1B_1B_4$. Via comparison of the reduced sequence, it will be easier to understand the biological function of various kinds of amino acid residues.

2.2. Graphical Representation of Protein Sequences. In this section, we will construct DV-curve representation of protein sequence. Given any protein primary sequence with length n, we can transform it into a new sequence composed of a character set of B_1, B_2, B_3, B_4. As shown in Figure 1, these alphabets are assigned, respectively, by consecutive vectors as follows:

$$\begin{aligned} B_1 &\Longrightarrow (1,1),(1,1) \\ B_2 &\Longrightarrow (1,1),(1,-1) \\ B_3 &\Longrightarrow (1,-1),(1,1) \\ B_4 &\Longrightarrow (1,-1),(1,-1). \end{aligned} \tag{2}$$

We connect adjacent dots with lines and then obtain a dual-vector curve form. This process is shown in Figure 2.

Based on the construction of DV-curve, we obtain two mathematical models, respectively. One is "from protein sequence to DV-curve," and the other is "from DV-curve to protein sequence." Firstly, we give some common symbols and variables. (1) According to the classification rule, we describe a protein sequence as $G(S) = b_1b_2b_3\cdots b_n$, where $b_i \in \{B_1, B_2, B_3, B_4\}$ with length n. It means that the protein sequence S is connected by these alphabets. (2) (x_i, y_i) is the coordinate of the ith point of DV-curve, and $(x_0, y_0) = (0, 0)$ is the start point.

Model One. Given a primary protein sequence, we can draw its DV-curve:

$$\begin{aligned} x_{2i-1} &= 2i-1, \quad i = 1, 2, \ldots, n, \\ x_{2i} &= 2i, \quad i = 1, 2, \ldots, n, \\ y_{2i-1} &= \begin{cases} y_{2i-2} + 1, & \text{if } b_i = B_1 \text{ or } B_2, \\ y_{2i-2} - 1, & \text{if } b_i = B_3 \text{ or } B_4, \end{cases} \\ y_{2i} &= \begin{cases} y_{2i-1} + 1, & \text{if } b_i = B_1 \text{ or } B_3, \\ y_{2i-1} - 1, & \text{if } b_i = B_2 \text{ or } B_4. \end{cases} \end{aligned} \tag{3}$$

According to the above four formulas, the coordinate of each point (x_i, y_i) can be calculated. Then we connect all the points with beelines, and the DV-curve is obtained.

Model Two. Given a DV-curve, we can also obtain the coarse-grained description of the protein sequence based on the detailed HP-model:

$$G(S_i) = \begin{cases} B_1, & \text{if } y_{2i-1} - y_{2i-2} = 1, \quad y_{2i} - y_{2i-1} = 1, \\ B_2, & \text{if } y_{2i-1} - y_{2i-2} = 1, \quad y_{2i} - y_{2i-1} = -1, \\ B_3, & \text{if } y_{2i-1} - y_{2i-2} = -1, \quad y_{2i} - y_{2i-1} = 1, \\ B_4, & \text{if } y_{2i-1} - y_{2i-2} = -1, \quad y_{2i} - y_{2i-1} = -1. \end{cases} \tag{4}$$

Here $i = 1, 2, 3, \ldots, n$. If each point (x_i, y_i) of DV-curve is given in this model, we can get each B_i according to the above formulas. So the simplified protein sequence $G(S) = b_1b_2\cdots b_n$ can be recovered; here $b_i \in \{B_1, B_2, B_3, B_4\}$ with length n.

3. Numerical Characterization of Protein Sequences

In order to facilitate quantitative comparisons of sequences, we will give numerical characterization of graphical curve as

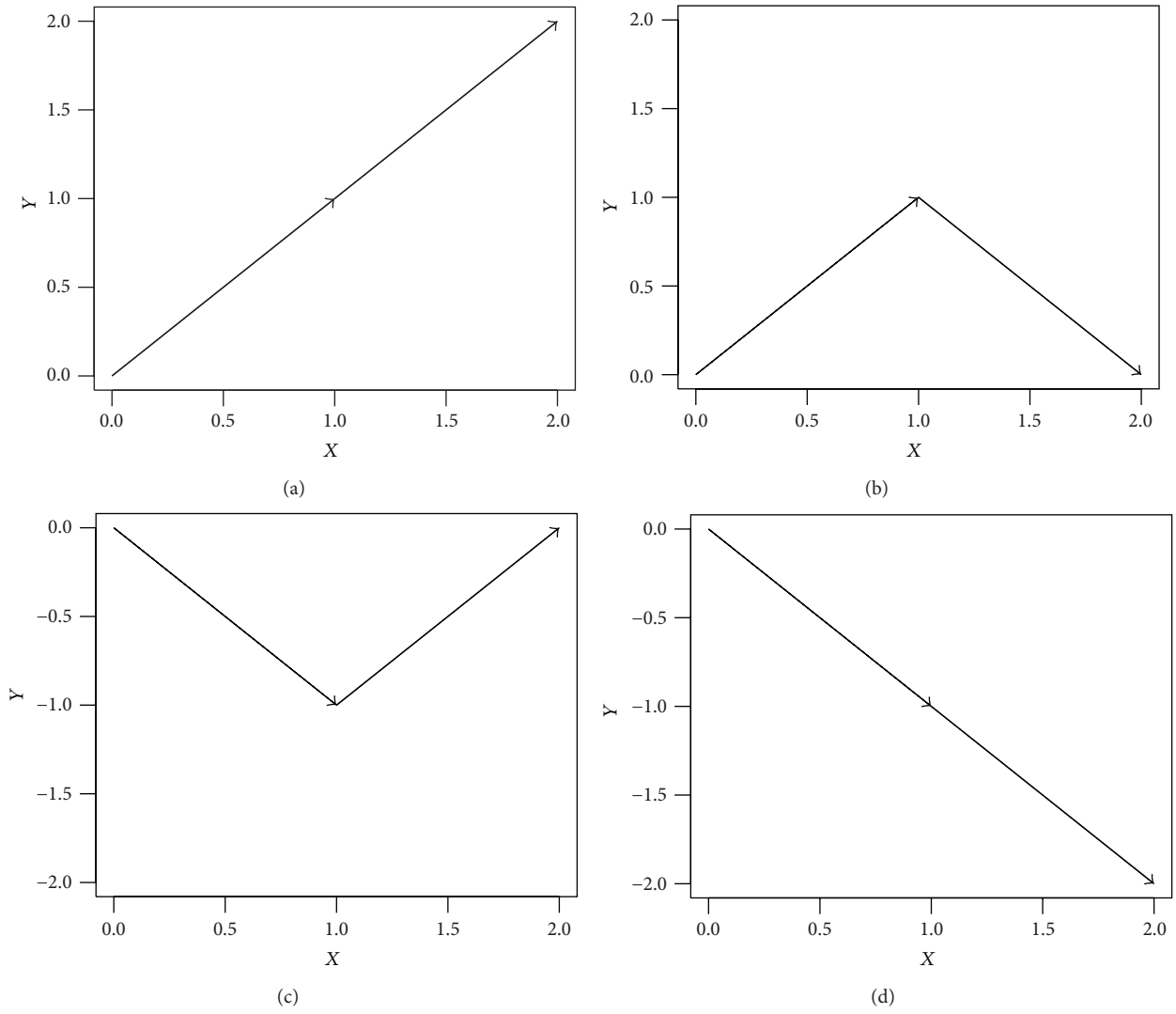


Figure 1: The representation of four alphabets of DV-curve: (a) B_1, (b) B_2, (c) B_3, and (d) B_4.

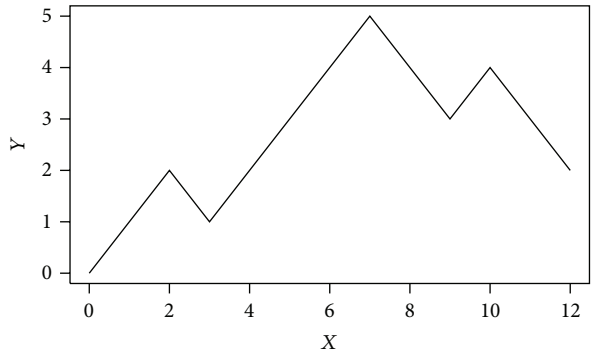


Figure 2: The DV-curve of sequence "WTFESR."

the descriptor. In general, we transform the graphical representation into a mathematical object like a matrix in order to draw some invariants. The frequently used matrices include E matrix, M matrix, L matrix, and L^k matrix proposed by Randić et al. [6, 8, 32–34]. Of course, there are some other matrix invariants such as the average matrix element, the average row sum, the Wiener number, and the ALE-index et al. These methods were used widely and proved to be useful. Here, we use the CM_{xy} as an alternative sequence invariant proposed by Liao et al. [35]:

$$(x_c, y_c) = \left(\frac{1}{2n+1} \sum_{i=0}^{2n} x_i,\ \frac{1}{2n+1} \sum_{i=0}^{2n} y_i \right),$$
$$CM_{xy} = \frac{1}{2n+1} \sum_{i=0}^{2n} (x_i - x_c)(y_i - y_c). \quad (5)$$

Obviously, this index is relatively simple for calculation so that this index can provide some convenience for long sequences.

If we adjust the order of B_1, B_2, B_3, B_4 corresponding to basic dual vectors, we can get another curve. So for a given sequence, we can get 4! = 24 different DV-curves totally. Therefore, a protein primary sequence can

Table 1: The information of 35 coronavirus spike proteins.

Number	Accession number	Abbreviation notation	Length (aa)	Group
1	P10033	FCoV1	1452	I
2	Q66928	FCoV2	1454	I
3	Q91AV1	PEDV3	1383	I
4	Q9DY22	TGEV4	1449	I
5	P18450	TGEV5	1449	I
6	P36300	CCoV6	1451	I
7	Q9J3E7	MHV7	1324	II
8	Q83331	MHV8	1361	II
9	P11224	MHV9	1324	II
10	O55253	MHV10	1360	II
11	Q9IKD1	RtCoV11	1360	II
12	P25190	BCoV12	1363	II
13	P15777	BCoV13	1363	II
14	Q9QAR5	BCoV14	1363	II
15	P36334	BCoV15	1363	II
16	P36334	HCoV16	1353	II
17	Q82666	IBV17	1166	III
18	P05135	IBV18	1163	III
19	P12722	IBV19	1154	III
20	Q64930	IBV20	1168	III
21	Q82624	IBV21	1159	III
22	P11223	IBV22	1162	III
23	Q98Y27	IBV23	1162	III
24	AAP41037	SCoV24	1255	IV
25	AAP300030	SCoV25	1255	IV
26	AAR91586	SCoV26	1255	IV
27	AAP51227	SCoV27	1255	IV
28	AAP33697	SCoV28	1255	IV
29	AAP13441	SCoV29	1255	IV
30	AAQ01597	SCoV30	1255	IV
31	AAU81608	SCoV31	1255	IV
32	AAS00003	SCoV32	1255	IV
33	AAR86788	SCoV33	1255	IV
34	AAR23250	SCoV34	1255	IV
35	AAT76147	SCoV35	1255	IV

be characterized by a 24-component vector as follows: $\vec{v} = [CM1_{xy}, CM2_{xy}, \ldots, CM24_{xy}]$. Based on the vectors, we can compare different protein sequences. Generally speaking, we can obtain the similarities of the two vectors by calculating Euclidean distance. If two sequences are similar, the distance between two corresponding points should be small. Given two species i and j, the corresponding vectors are $\vec{v}_i = [CMi1_{xy}, CMi2_{xy}, \ldots, CMi24_{xy}]$ and $\vec{v}_j = [CMj1_{xy}, CMj2_{xy}, \ldots, CMj24_{xy}]$, respectively; then we have $d(\vec{v}_i, \vec{v}_j) = \sqrt{\sum_{k=1}^{24}(CMik_{xy} - CMjk_{xy})^2}$.

4. Application

The comparison on biology sequences is one of the most important parts in bioinformatics when analyzing similarities of function and properties. In this section, we will give two main applications of this new graphical representation. One is similarity analysis based on visual graphics. Generally, similarity analysis can be divided into two types of methodologies to conduct the comparison: sequence alignment and sequence descriptors comparison. When recognizing figures, our brain is more helpful for similarity analysis in multiple sequences. So it is desirable to propose similarity analysis by inspecting the DV-curve of protein. The other is evolutionary homology analysis based on the numerical characterization of DV-curve, and we construct a 24-component vector to characterize any protein sequence. As further work, the phylogenetic tree of 35 coronavirus spike proteins is constructed.

4.1. Similarity Analysis Based on Visual Inspection of the Protein DV-Curve Graphs. Since Smith and Waterman developed a dynamic programming algorithm in 1981, many alignment algorithms identifying whether two biological sequences are similar to each other have been studied. These methods are proved to be efficient. However, multiple sequence alignment (MSA) of several hundred sequences has always produced a bottleneck.

In 1994, MSA was proved to be an NP-complete problem by Wang and Jiang [36]. Moreover, most experts think that it is impossible until now to build a deterministic polynomial algorithm to handle an NP-complete problem. It needs to exhaust almost billions or trillions of years. Except long computational time, there also exists possible bias of multiple sequence alignments for multiple occurrences of highly similar sequence [37].

However, our brain is much more powerful than computer when recognizing different figures. So it can help us to analyze the similarity in multiple sequences. If we can provide a simple, intuitional, clear, and nondegenerate 2D graphical representation of protein sequences, molecular biologists may easily find out which sequence is most similar or dissimilar to the given target sequence. And next they can use alignment algorithms for further confirmation.

According to our proposed definition of protein DV-curve, we can draw the curves of some ND6 (NADH dehydrogenase subunit 6) proteins in order to conveniently compare them. Protein sequences that are used to prove our approach were downloaded from GenBank: human (*YP*_003024037.1), gorilla (*NP*_008223), chimpanzee (*NP*_008197), wallaroo (*NP*_007405), harbor seal (H. seal) (*NP*_006939), gray seal (G. seal) (*NP*_007080), rat (*AP*_004903), and mouse (*NP*_904339), and the same data set was used in [26, 27].

In Figure 3, it is evident that protein graph of wallaroo is obviously different from the other species because it is the most remote species from the remaining mammals. Furthermore, we can see human and chimpanzee have similar curves, harbor seal and gray seal's curves are almost identical, and two curves of rat and mouse are very similar. All these results not only are consistent with the conclusions drawn by Smith-Waterman algorithm, but also agree well with the known fact of evolution and results drawn by other authors [26, 27, 38–40]. In particular, compared with the conclusion of [27], the DV-curve representation reflecting the similarities of sequences is more simple, intuitional, and visible.

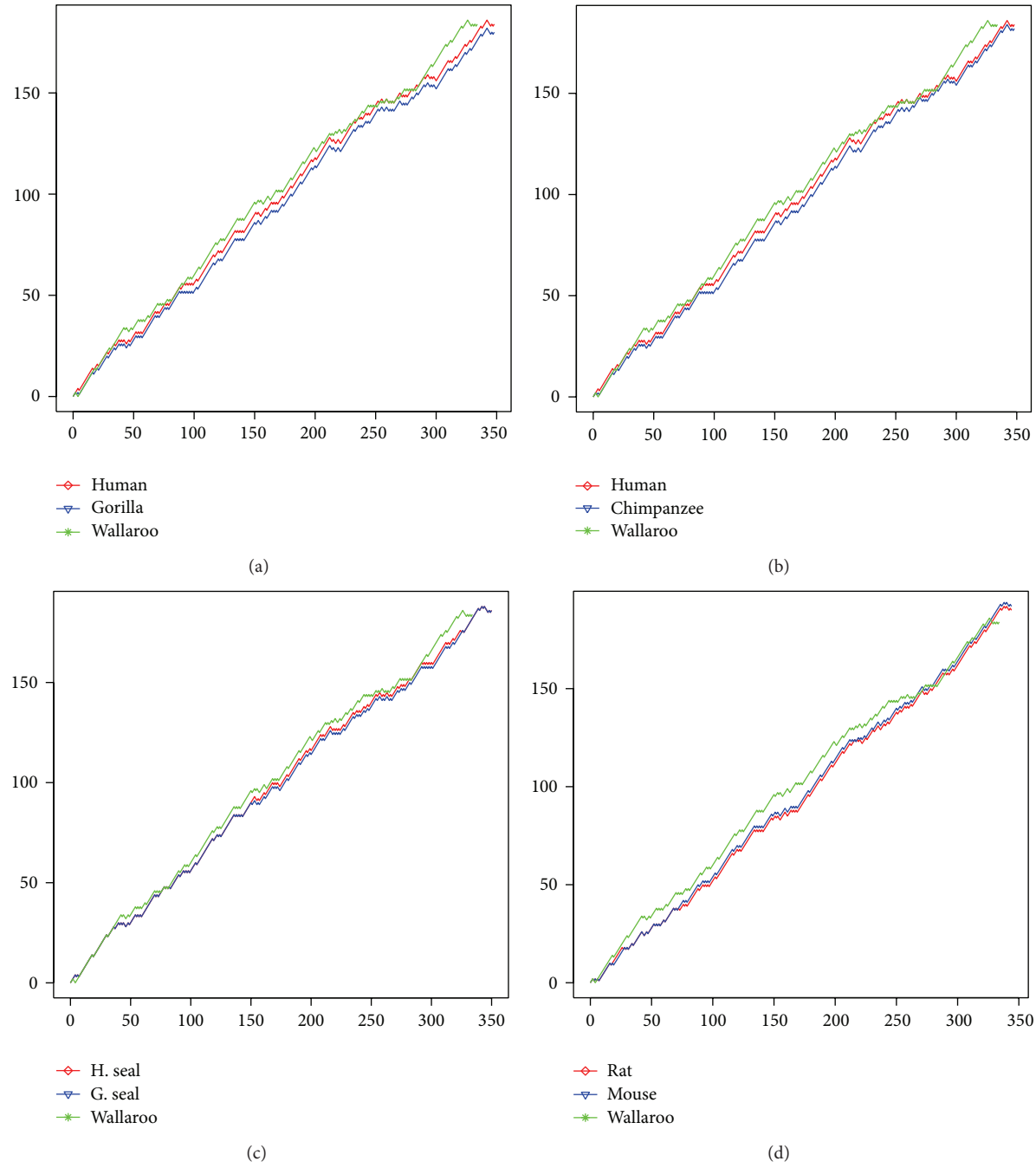


Figure 3: The DV-curve graphical representations of different ND6 proteins.

4.2. The Phylogenetic Analysis among the Spike Glycoprotein of Coronaviruses. Coronaviruses belong to order Nidovirales, family Coronaviridae, and genus *Coronavirus*. They are a diverse group of large, enveloped, single-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals. Generally, coronaviruses can be divided into three groups: the first group and the second group come from mammalian; the third group comes from poultry (chicken and turkey). A novel coronavirus has been identified as the cause of the outbreak of severe acute respiratory syndrome (SARS). Previous phylogenetic analysis based on sequence alignments shows that SARS-CoVs come from a new group distantly related to the above three groups of previously characterized coronaviruses [41, 42]. The spike (S) protein, which is common to all known coronaviruses, is crucial for viral attachment and entry into the host cell. To illustrate the use of DV-curve of protein sequences, we will construct the phylogenetic tree of 35 coronavirus spike proteins of Table 1.

As we have described above, a protein sequence can be associated with a 24-component vector. Given two species i and j, we can calculate the distance between them. Our

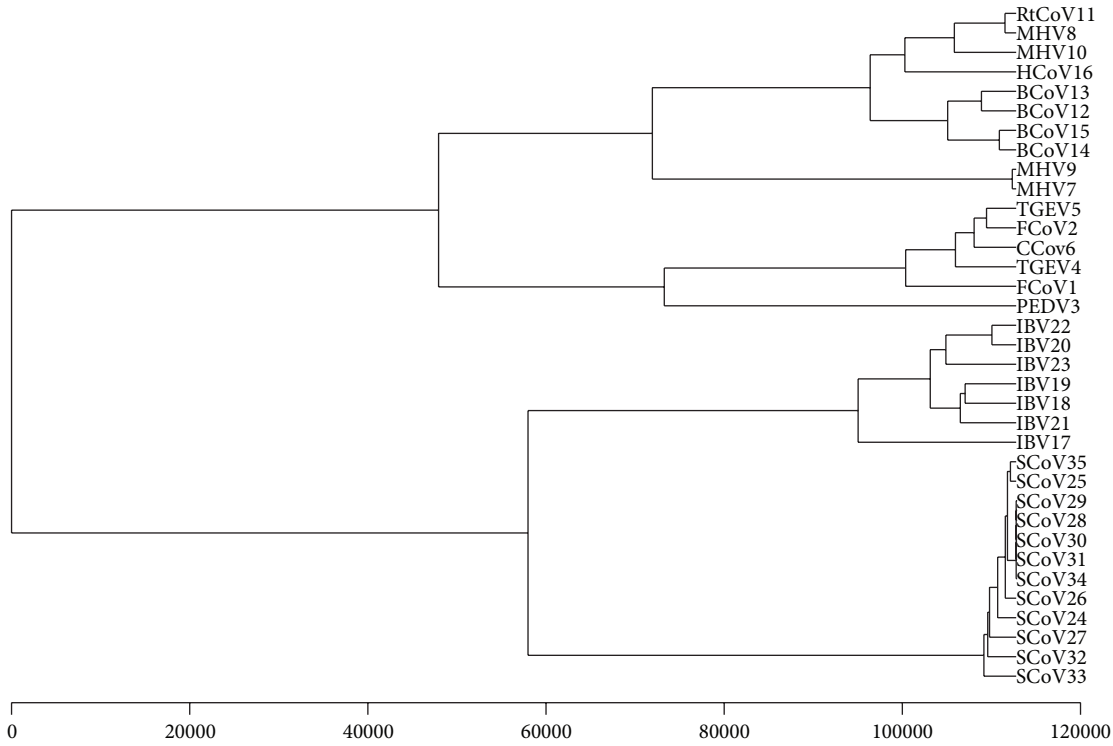


Figure 4: The phylogenetic tree based on the spike proteins.

datasets used in this paper were downloaded from GenBank (see Table 1 for details). Corresponding to 35 spike proteins, a 35×35 real symmetric matrix $D = (d_{ij})$ is obtained and used to reflect the evolutionary distance of them. Using the UPGMA program included in PHYLIP package 3.65, we can construct the phylogenetic tree of these 35 species [43, 44]. The branch lengths are not scaled according to the distances and only the topology of the tree is concerned.

Figure 4 shows coronaviruses can be overall divided into four groups. Furthermore, it is evident that SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from the other three groups of coronaviruses.

RtCoV11, MHV8, MHV10, HCoV16, BCoV13, BCoV12, BCoV15, BCoV14, MHV9, and MHV7, which belong to group 2, are situated at an independent branch, while TGEV5, FCoV2, CCoV6, TGEV4, FCoV1, and PEDV3, belonging to group 1, tend to cluster together. Meanwhile, the group 3 coronaviruses, including IBV22, IBV20, IBV23, IBV19, IBV18, IBV21, and IBV17, tend to cluster together in another branch. The resulting monophyletic clusters agree well with the established taxonomic groups [45, 46]. The conclusion is similar to that reported by other authors [23, 24]. Compared with result [24], it is noteworthy that a closer look at the subtree of the first branch shows coronavirus from three different species; that is, MHV, BCoV, and HCoV can be separated clearly, while they cluster together in a subtree by Li's method. Obviously, our conclusion is more consistent with the known evolution fact.

5. Conclusion

According to the detailed hydrophobic-hydrophilic (HP) model of amino acids, we can reduce a protein primary sequence containing 20 amino acids into a four-letter sequence, which can be treated as a coarse-grained description of the protein primary sequence. Here we cannot avoid losing some information in the reduced sequences, but we can focus our main attention on the part of our interest.

Some alignment-free methods to analyze DNA sequences have been proposed. However, there are few alignment-free methods to analyze protein sequences. Our method realizes the generalization from DNA graphical representations to those of proteins acceptable and can be seen a valid supplement to graphical representation of protein sequences. Meanwhile we first propose to combine DV-curve and the detailed HP model together to describe protein sequences.

Compared with classical Smith-Waterman algorithm, the similarity/dissimilarity analysis results are consistent with DV-curve. In addition, the advantage of our method is that it can visualize the local and global features among different proteins no matter how long these sequences are and avoid degeneracy at the same time. The new approach is applied in two aspects: one is similarity intuitive analysis of ND6 protein sequences of several species and the other is phylogenetic analysis among 35 coronaviruses based on their spike proteins. Results have shown that our proposed method is more intuitional, simple, effectual, and feasible.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank to all the anonymous reviewers for their valuable suggestions and support. This research is supported by the National Science Foundation of China Grants 11371227 and 10921101.

References

[1] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.

[2] M. A. Gates, "A simple way to look at DNA," *Journal of Theoretical Biology*, vol. 119, no. 3, pp. 319–328, 1986.

[3] A. Nandy, "Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences," *Computer Applications in the Biosciences*, vol. 12, no. 1, pp. 55–62, 1996.

[4] X. F. Guo, M. Randic, and S. C. Basak, "A novel 2-D graphical representation of DNA sequences of low degeneracy," *Chemical Physics Letters*, vol. 350, no. 1-2, pp. 106–112, 2001.

[5] A. Nandy and P. Nandy, "On the uniqueness of quantitative DNA difference descriptions in 2D graphical representation models," *Chemical Physics Letters*, vol. 368, no. 1-2, pp. 102–107, 2003.

[6] M. Randic, M. Vracko, N. Lers, and D. Plavsic, "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation," *Chemical Physics Letters*, vol. 371, pp. 202–207, 2003.

[7] Y. H. Yao and T.-M. Wang, "A class of new 2-D graphical represent ation of DNA sequences and their application," *Chemical Physics Letters*, vol. 398, no. 4–6, pp. 318–323, 2004.

[8] M. Randic, "Graphical representations of DNA as 2-D map," *Chemical Physics Letters*, vol. 386, pp. 468–471, 2004.

[9] G. H. Huang, B. Liao, Y. F. Li, and Z. B. Liu, "H-L curve: a novel 2D graphical representation for DNA sequences," *Chemical Physics Letters*, vol. 462, no. 1–3, pp. 129–132, 2008.

[10] Z.-J. Zhang, "DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences," *Bioinformatics*, vol. 25, no. 9, pp. 1112–1117, 2009.

[11] W. Deng and Y. H. Luan, "Analysis of similarity/dissimilarity of DNA sequences based on chaos game representation," *Abstract and Applied Analysis*, vol. 2013, Article ID 926519, 6 pages, 2013.

[12] B. Liao and K. Ding, "A 3D graphical representation of DNA sequences and its application," *Theoretical Computer Science*, vol. 358, no. 1, pp. 56–64, 2006.

[13] Z. Cao, B. Liao, and R. Li, "A group of 3D graphical representation of DNA sequences based on dual nucleotides," *International Journal of Quantum Chemistry*, vol. 108, no. 9, pp. 1485–1490, 2008.

[14] Y. J. Huang and T. M. Wang, "New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis," *International Journal of Quantum Chemistry*, vol. 112, no. 6, pp. 1746–1757, 2012.

[15] B. Liao, Y. S. Zhang, K. Q. Ding, and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation," *Journal of Molecular Structure: THEOCHEM*, vol. 717, no. 1–3, pp. 199–203, 2005.

[16] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences," *Chemical Physics Letters*, vol. 407, no. 1–3, pp. 63–67, 2005.

[17] B. Liao, R. Li, W. J. Zhu, and X. Xiang, "On the similarity of DNA primary sequences based on 5-D representation," *Journal of Mathematical Chemistry*, vol. 42, no. 1, pp. 47–57, 2007.

[18] B. Liao and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1666–1670, 2004.

[19] M. Randić, J. Zupan, and A. T. Balaban, "Unique graphical representation of protein sequences based on nucleotide triplet codons," *Chemical Physics Letters*, vol. 397, no. 1-3, pp. 247–252, 2004.

[20] F. L. Bai and T. M. Wang, "A 2-D graphical representation of protein sequences based on nucleotide triplet codons," *Chemical Physics Letters*, vol. 413, no. 4–6, pp. 458–462, 2005.

[21] N. Liu and T. M. Wang, "Protein-based phylogenetic analysis by using hydropathy profile of amino acids," *FEBS Letters*, vol. 580, no. 22, pp. 5321–5327, 2006.

[22] M. Randić, "2-D Graphical representation of proteins based on physico-chemical properties of amino acids," *Chemical Physics Letters*, vol. 440, no. 4–6, pp. 291–295, 2007.

[23] C. Li, L. L. Xing, and X. Wang, "2-D graphical representation of protein sequences and its application to coronavirus phylogeny," *Journal of Biochemistry and Molecular Biology*, vol. 41, no. 3, pp. 217–222, 2008.

[24] D. D. Li, J. Wang, and C. Li, "New 3-D graphical representation of protein sequences and its application," *China Journal of Bioinformatics*, vol. 7, no. 1, pp. 60–63, 2009.

[25] J. Wen and Y. Zhang, "A 2D graphical representation of protein sequence and its numerical characterization," *Chemical Physics Letters*, vol. 476, no. 4–6, pp. 281–286, 2009.

[26] Y. H. Yao, Q. Li, N. Li, X. Y. Nan, P. A. He, and Y. Z. Zhang, "Similarity/dissimilarity studies of protein sequences based on a new 2d graphical representation," *Journal of Computational Chemistry*, vol. 31, no. 5, pp. 1045–1052, 2010.

[27] X.-L. Xie, L.-F. Zheng, Y. Yu et al., "New technique: protein sequence analysis based on hydropathy profile of amino acids," *Journal of Zhejiang University: Science B*, vol. 13, no. 2, pp. 152–158, 2012.

[28] M. I. Abo El Maaty, M. M. Abo-Elkhier, and M. A. Abd Elwahaab, "3D graphical representation of protein sequences and their statistical characterization," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 21, pp. 4668–4676, 2010.

[29] M. M. Abo-EIkhier, "Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor," *Journal of Biophysical Chemistry*, vol. 3, no. 2, pp. 142–148, 2012.

[30] J. Wang and W. Wang, "Modeling study on the validity of a possibly simplified representation of proteins," *Physical Review E*, vol. 61, no. 6, pp. 6981–6986, 2000.

[31] T. A. Brown, *Genetics*, Chapman & Hall, London, UK, 3rd edition, 1998.

[32] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, "On 3-D graphical representation of DNA primary sequences and their numerical characterization," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1235–1244, 2000.

[33] M. Randic, M. Vracko, L. Nelia, and P. Dejan, "Novel 2-D graphical representation of DNA sequences and their numerical characterization," *Chemical Physics Letters*, vol. 368, no. 1-2, pp. 1–6, 2003.

[34] M. Randic, M. Vracko, J. Zupan, and M. Novic, "Compact 2-D graphical representation of DNA," *Chemical Physics Letters*, vol. 373, pp. 558–562, 2003.

[35] B. Liao, M. Tan, and K. Ding, "Application of 2-D graphical representation of DNA sequence," *Chemical Physics Letters*, vol. 414, pp. 296–300, 2005.

[36] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment," *Journal of Computational Biology*, vol. 1, no. 4, pp. 337–348, 1994.

[37] T. D. Pham and J. Zuegg, "A probabilistic measure for alignment-free sequence comparison," *Bioinformatics*, vol. 20, no. 18, pp. 3455–3461, 2004.

[38] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[39] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[40] V. Makarenkov and F.-J. Lapointe, "A weighted least-squares approach for inferring phylogenies from incomplete distance matrices," *Bioinformatics*, vol. 20, no. 13, pp. 2113–2121, 2004.

[41] T. G. Ksiazek, S. R. Zaki, C. Urbani et al., "A novel coronavirus associated with severe acute respiratory syndrome," *The New England Journal of Medicine*, vol. 348, pp. 1953–1966, 2003.

[42] M. A. Marra, S. J. Jones, C. R. Astell et al., "The genome sequence of the sars-associated coronavirus," *Science*, vol. 300, p. 1399, 2003.

[43] P. H. A, R. R. Sneath, and Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, 1973.

[44] PHILIP, http://evolution.gs.washington.edu/phylip.html.

[45] 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.

[46] S. K. P. Lau, P. C. Y. Woo, K. S. M. Li et al., "Severe acute respiratorysyndrome coronavirus-like virus in Chinese horseshoe bats," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 14040–14045, 2005.