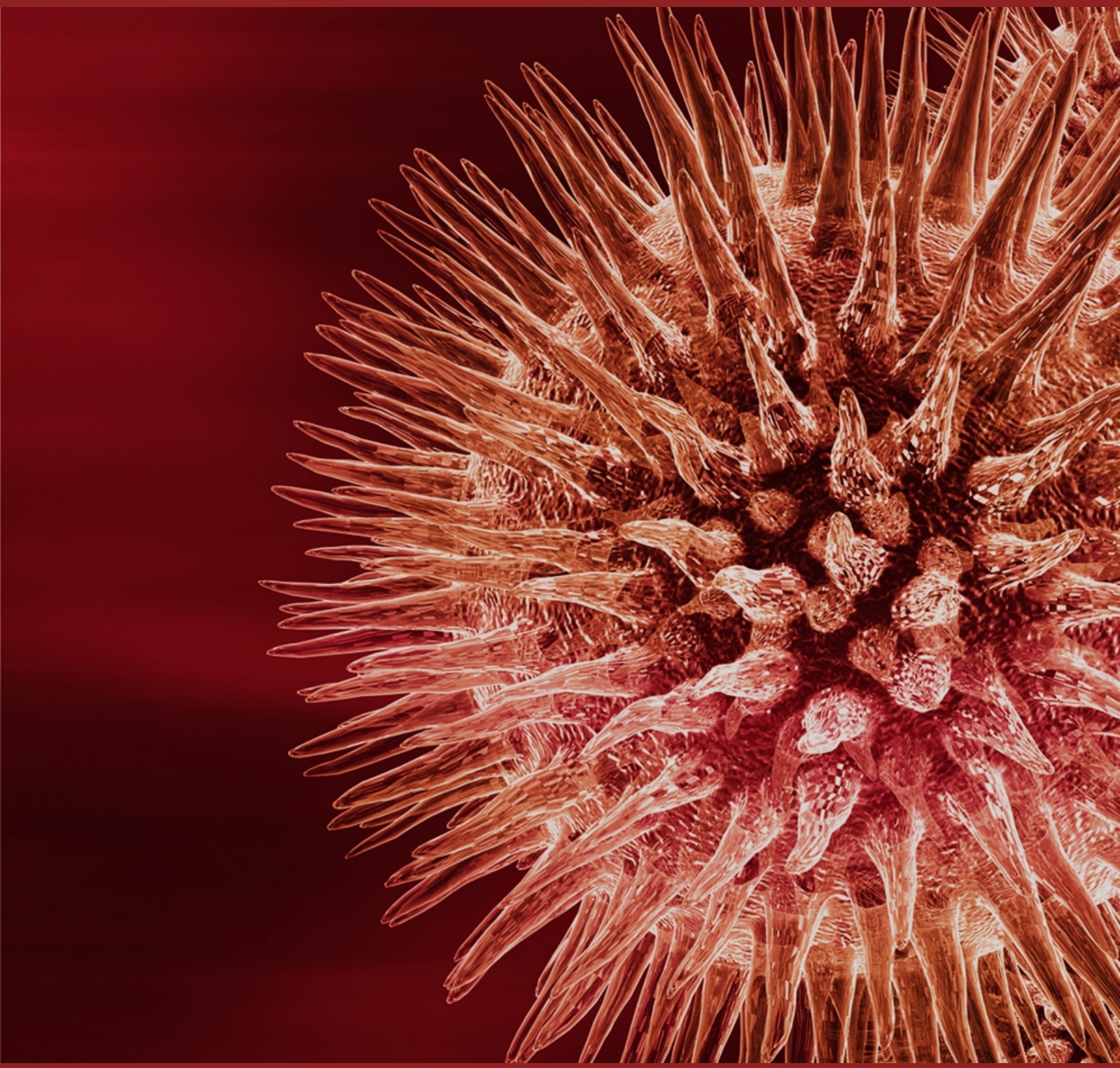
Journal of Biomedicine and Biotechnology

Data Mining in Genomics and Proteomics

Guest Editors: Halima Bensmail and Abdelali Haoudi



Journal of Biomedicine and Biotechnology

Data Mining in Genomics and Proteomics

Journal of Biomedicine and Biotechnology

Data Mining in Genomics and Proteomics

Guest Editors: Halima Bensmail and Abdelali Haoudi

Contents

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 63–64 • DOI: 10.1155/JBB.2005.63

EDITORIAL

Data Mining in Genomics and Proteomics

Halima Bensmail[1] and Abdelali Haoudi[2]

[1]*Department of Statistics, The University of Tennessee, Knoxville, TN 37996-0532, USA*
[2]*Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23507-1696, USA*

There is no doubt that both computational biology and bioinformatics, and the interface of computer science and biology in general, are central to the future of biological research. The disciplines span a process that begins with data collection, analysis, classification, and integration, and ends with interpretation, modeling, visualization, and prediction. Data mining plays a role in the middle of this process. Overall, the focus is on identifying opportunities and developing computational solutions (including algorithms, models, tools, and databases) that can be used for experimental design, data analysis and interpretation, and hypothesis generation.

Data mining is the search for hidden trends within large sets of data. Data mining approaches are needed at all levels of genomics and proteomics analyses. These studies can provide a wealth of information and rapidly generate large quantities of data from the analysis of biological specimens from healthy and diseased tissues. The high dimensionality of data generated from these studies will require the development of improved bioinformatics and computational biology tools for efficient and accurate data analyses.

This issue of the Journal of Biomedicine and Biotechnology consists of seventeen papers that describe different applications of data mining to both genomics and proteomics studies in yeast, and plant and human cells and tissues. Papers by Bensmail et al, Ghosh and Chinnaiyan, and Mao et al present different classification and clustering approaches for disease biomarkers discovery. Genomics and proteomics studies have shown great promises and have been applied to studies aiming at generating expression profiles and elucidating expression networks in different organisms as shown in the papers by Samsa et al, Mungur et al, Liu et al, Baldwin et al, and Joy et al. Data mining in genomics and proteomics studies reveals new regulatory pathways and mechanisms in different health and disease conditions as presented by Wren and Garner, and provides comparative sequence analysis approaches as presented by Gambin and Otto and Gao et al. Those studies have also provided approaches for subcellular localization of proteins suggesting that such approaches can produce an objective systematics for protein location and provide an important starting point for discovering sequence motifs that determine localization as presented by Chen and Murphy. Chen et al studied the performance of five nonparameteric tests to select genes and proved that the popular F test does not perform well on gene expression data since the heterogeneity behavior assumption is the most dominant in the gene expression data. Corder et al explored a statistical approach called grade of membership (GOM) and proved that brain hypoperfusion contributes to dementia, possibly to Alzheimer's disease (AD) pathogenesis, and raises the possibility that the APOE ε^4 allele contributes directly to heart value and myocardial damage. Hand and Heard present in their review article various tools for finding relevant subgroups in gene expression data. Alkharouf et al conduct an OLAP cube (online analytical processing) to mine a time series experiment designed to identify genes associated with resistance of soybean to the soybean cyst nematode, which is a devastating pest of soybean. Brylinski et al created a sequence-to-structure library based on the complete PDB database. Then an early-stage folding conformation and information entropy were used for structure analysis and classification.

Whilst postgenomic science is producing vast data torrents, it is well known that data do not equal knowledge and so the extraction of the most meaningful parts of these data is key to the generation of useful new knowledge. More sophisticated data mining strategies are needed for mining such high-dimensional data to generate useful relationships, rules, and predictions.

Correspondence and reprint requests to Abdelali Haoudi, Eastern Virginia Medical School, Department of Microbiology and Molecular Cell Biology, Norfolk, VA 23507-1696, USA, E-mail: haoudia@evms.edu

Halima Bensmail
Abdelali Haoudi

Halima Bensmail received her PhD degree jointly from Pierre & Marie Curie University, Paris, France, and the University of Washington, Seattle, in 1996 in statistics and mathematical modelling. She then joined the University of Washington for a Researcher position for a period of two years, then worked as a Consultant and Associate Researcher at the Fred Hutchinson Cancer Research Center. She joined the University of Leiden for a period of three years. She joined the University of Tennessee for a Statistics Assistant Professor position at the Department of Statistics in 2000. Dr. Bensmail is also an Associate Editor for the Journal of Biomedicine and Biotechnology and a Reviewer at the NIH and NSF. She has established numerous collaborations both within the academia (EVMS, Georgia State University, University of California, Oak Ridge Laboratory) and with the private sector (HIRST Company for Hedge Fund Strategy Benchmarks, Federal Bank of Atlanta). She has advised numerous PhD and Master's students and cochaired many conferences particularly on data mining. She is a member of several scientific organizations and has received numerous scientific and teaching awards.

Abdelali Haoudi received his PhD degree in cellular and molecular genetics jointly from Pierre & Marie Curie University and Orsay University in Paris, France. He then joined the National Institutes of Health (NIEHS, NIH) for a period of four years after winning the competitive and prestigious NIH Fogarty International Award. Dr. Haoudi then joined the faculty in the Department of Microbiology and Molecular Cell Biology at Eastern Virginia Medical School in Norfolk, Va, in 2001. Dr. Haoudi is primarily interested in the elucidation of the molecular basis of cancer including cell cycle checkpoints, DNA repair and apoptosis, in addition to the development of cancer gene therapeutic strategies. Dr. Haoudi is also the codirector of the Cancer Biology and Virology Focal Group. He has founded the Journal of Biomedicine and Biotechnology (http://www.hindawi.com/journals/jbb) and is also the Founder and President of the International Council of Biomedicine and Biotechnology (http://www.i-council-biomed-biotech.org). He is a member of several scientific organizations and has received numerous scientific awards.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 65–79 • DOI: 10.1155/JBB.2005.65

RESEARCH ARTICLE

Early-Stage Folding in Proteins *(In Silico)* Sequence-to-Structure Relation

Michał Brylinski,[1,2] Leszek Konieczny,[3] Patryk Czerwonko,[1] Wiktor Jurkowski,[1,2] and Irena Roterman[1]

[1] *Department of Bioinformatics and Telemedicine, Medical Faculty, Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland*
[2] *Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland*
[3] *Institute of Biochemistry, Medical Faculty, Jagiellonian University, Kopernika 7, 31-501 Cracow, Poland*

Received 16 September 2004; revised 3 January 2005; accepted 5 January 2005

A sequence-to-structure library has been created based on the complete PDB database. The tetrapeptide was selected as a unit representing a well-defined structural motif. Seven structural forms were introduced for structure classification. The early-stage folding conformations were used as the objects for structure analysis and classification. The degree of determinability was estimated for the sequence-to-structure and structure-to-sequence relations. Probability calculus and informational entropy were applied for quantitative estimation of the mutual relation between them. The structural motifs representing different forms of loops and bends were found to favor particular sequences in structure-to-sequence analysis.

INTRODUCTION

Prediction of three-dimensional protein structures remains a major challenge to modern molecular biology. On the one hand, identical pentapeptide sequences exist in completely different tertiary structures in proteins [1]; on the other, different amino acid sequences can adopt approximately the same three-dimensional structure. However, the patterns of sequence conservation can be used for protein structure prediction [2, 3, 4]. Usually, secondary structure definition has been used for ab initio methods as a common starting conformation for protein structure prediction [5]. A large body of experiments and theoretical evidence suggests that local structure is frequently encoded in short segments of protein sequence. A definite relation between the amino acid sequences of a region folded into a supersecondary structure has been found. It was also found that they are independent of the remaining sequence of the molecule [6, 7]. Early studies of local sequence-structure relationships and secondary structure prediction were based on either simple physical principles [8] or statistics [9, 10, 11, 12]. Nearest-neighbor methods use a database of proteins with known three-dimensional structures to predict the conformational states of test protein [13, 14, 15, 16]. Some methods are based on nonlinear algorithms known as neural nets [17, 18, 19] or hidden Markov models [20, 21, 22, 23]. In addition to studies of sequence-to-structure relationships focused on determining the propensity of amino acids for predefined local structures [24, 25, 26, 27], others involve determining patterns of sequence-to-structure correlations [21, 22, 28, 29, 30]. The evolutionary information contained in multiple sequence alignments has been widely used for secondary structure prediction [31, 32, 33, 34, 35, 36, 37, 38]. Prediction of the percentage composition of α-helix, β-strand, and irregular structure based on the percentage of amino acid composition, without regard to sequence, permits proteins to be assigned to groups, as all α, all β, and mixed α/β [5, 39].

Structure representation is simplified in many models. Side chains are limited to one representative virtual atom; virtual $C\alpha - C\alpha$ bonds are often introduced to decrease the number of atoms present in the peptide bond [40, 41]. The search for structure representation in other than the ϕ, ψ angles conformational space has been continuing [42].

Other models are based on limitation of the conformational space. One of them divided the Ramachandran map into four low-energy basins [43, 44]. In another study, all sterically allowed conformations for short polyalanine chains were enumerated using discrete bins

Correspondence and reprint requests to Irena Roterman, Department of Bioinformatics and Telemedicine, Medical Faculty, Jagiellonian University, Kopernika 17, 31-501, Poland, E-mail: myroterm@cyf-kr.edu.pl

called mesostates [45]. The need to limit the confomational space was also asserted [46, 47].

The model introduced in this paper is based on limitation of the conformational space to the particular part of the Ramachandran map. The structures created according to this limited conformational subspace are assumed to represent early-stage structural forms of protein folding *in silico*.

In this paper, in contrast to commonly used base of final native structures of proteins, the early-stage folding conformation of the polypeptide chain is the criterion for structure classification.

Two approaches are the basis for the early-stage folding model presented in this paper.

(1) The geometry of the polypeptide chain can be expressed using parameters other than ϕ, ψ angles. These new parameters are the V-angle—dihedral angle between two sequential peptide bond planes—and the R-radius, radius of curvature, found to be dependent on the V-angle in the form of a second-degree polynomial. Details on the background of the geometric model based on the V, R [48, 49] are recapitulated briefly in "appendix A."

(2) The structures satisfying the V-to-R relation appeared to distinguish the part of the Ramachandran map (the complete conformational space) delivering the limited conformational subspace (ellipse path on the Ramachandran map). It was shown that the amount of information carried by the amino acid is significantly lower than the amount of information needed to predict ϕ, ψ angles (point on Ramachandran map). These two amounts of information can be balanced after introducing the conformational subspace limited to the conformational subspace distinguished by the simplified model presented above. Details on the background of the information-theory-based model [50] are reviewed briefly in "appendix B."

The conformational subspace found to satisfy the geometric characteristics (polypeptide limited to the chain peptide bond planes with side chains ignored) and the condition of information balancing appeared to select the part of Ramachandran map which can be treated as the early-stage conformational subspace.

The introduced model of early-stage folding was extended to make it applicable to the creation of starting structural forms of proteins for an energy-minimization procedure oriented to protein structure prediction. The characteristics and possible applicability of the sequence-to-structure and structure-to-sequence contingency tables is the aim of this paper.

The structures created according to the limited conformational subspace can be reached in two different ways: (1) as the partial unfolding (Figures 1a–1e) and (2) as the basis for the initial structure assumed to represent early-stage folding (Figures 1f–1j). The partial unfolding of the native structural form (called the "step-back" structure in this paper) is expressed by changing the ϕ, ψ angles to the ϕ_{sb}, ψ_{sb} angles (ϕ_{sb}, ψ_{sb} angles belong to the ellipse path, and their values are obtained according to the criterion of the shortest distance between ϕ, ψ and the ellipse—shown in Figure 1b). The second approach, in which the structure is created on the basis of the ϕ_{es}, ψ_{es} angles (ϕ_{es}, ψ_{es} denote the dihedral angles belonging to the ellipse and representing a particular probability maximum), is based on the library of sequence-to-structure relations for tetrapeptides.

A scheme summarizing the two procedures—partial unfolding and partial folding—is shown below (Figure 1). The procedure called partial unfolding starts at the native structure of the protein (Figure 1a). The values of the ϕ, ψ angles present in the protein are changed (according to the shortest distance criterion) to the values of the angles belonging to the ellipse (ϕ_{sb}, ψ_{sb}). When these dihedral angles are applied, the structure of the same protein looks as is shown in Figure 1c. When this procedure is applied to all proteins present in the protein data bank, a probability profile can be obtained which represents the distribution of ϕ, ψ angles in the limited conformational subspace. The distribution is different for each amino acid, although some characteristic maxima can be distinguished. The profile shown in Figure 1d represents Glu (the ellipse equation t-parameter = 0° represents the point of $\phi = 90°$ and $\psi = -90°$, and then increases clockwise). Particular probability maxima can be recognized using the letter codes also shown in Figure 2. These letter codes are used to classify the structures of proteins in their early-stage folding (*in silico*) (Figure 1e).

The opposite procedure, aimed at protein folding, is shown also in Figures 1f–1j. The starting point in this procedure is the amino acid sequence of a particular protein. After selecting four-amino-acid fragments (in an overlapping system), four different structural codes (for the same tetrapeptide) can be attributed on the basis of the contingency table described above (Figure 1f). Only a particular fragment of the probability profile (according to the letter code) can be recognized in this case. In consequence, the ϕ_{es}, ψ_{es} values representing the location of the probability maximum on the t-axis can be attributed to a particular sequence (Figure 1g). This is why the ϕ_{es}, ψ_{es} angles differ versus ϕ_{sb}, ψ_{sb}. In consequence, the structure of the transforming growth factor β binding protein-like domain (protein selected as an example, PDB ID: 1APJ) created according to the ϕ_{es}, ψ_{es} angles shown in Figure 1h differs versus the (ϕ_{sb}, ψ_{sb})-based structure. The "sb" (step-back) and "es" (early-stage) structures differ due to the continuous form of the probability distribution in "sb" procedure and the discrete one in the "es" procedure. The next step in the prediction procedure is energy minimization, which in some cases causes approach toward the native structure (Figure 1j).

The structures created according to the ellipse path treated as the starting structures for the energy-minimization procedure, deliver forms that approach the native structure after one simple optimization procedure. BPTI [51], ribonuclease [50], to some extent also human hemoglobin α and β chains [52] and lysozyme [53] were used as the model molecules. All these examples

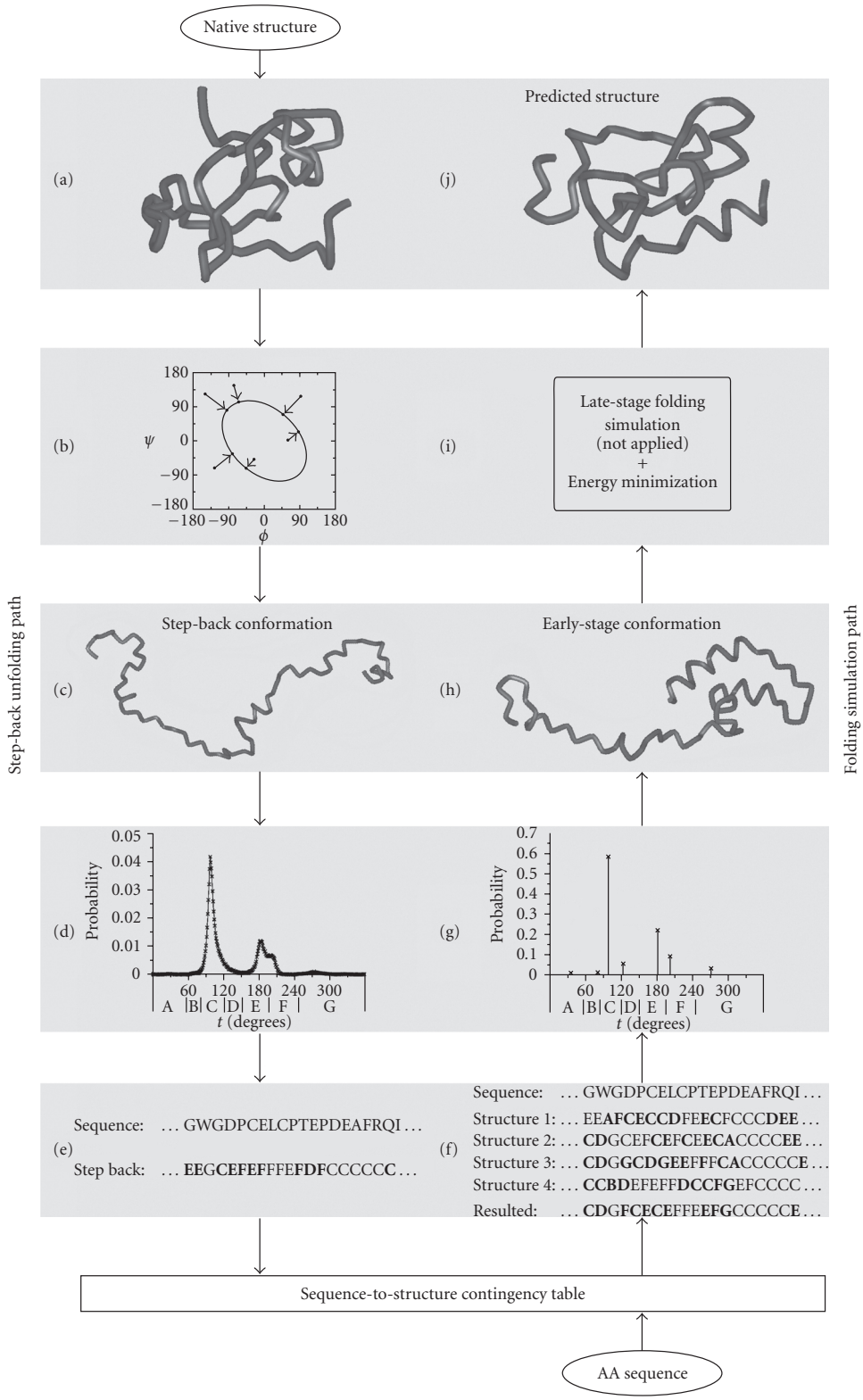


FIGURE 1. (a–e) Step-back unfolding path: (a) native structure of 1APJ, (b) partial unfolding procedure, (c) step-back conformation according to the limited conformational subspace, (d) example of amino-acid-dependent probability profile (Glu) for complete PDB 2003 after moving ϕ, ψ angles to the nearest point on the ellipse path, (e) letter codes assigned according to probability profiles. (f–j) Folding simulation path: (f) early-stage structure prediction in terms of structural letters, (g) an example of a discrete profile (Glu) applied to early-stage structure creation, (h) predicted early-stage conformation of 1APJ, (i) late-stage folding simulation procedure (under consideration—not applied yet), (j) structure of 1APJ as a result of the energy-minimization procedure with proper disulphide bridges constraints.

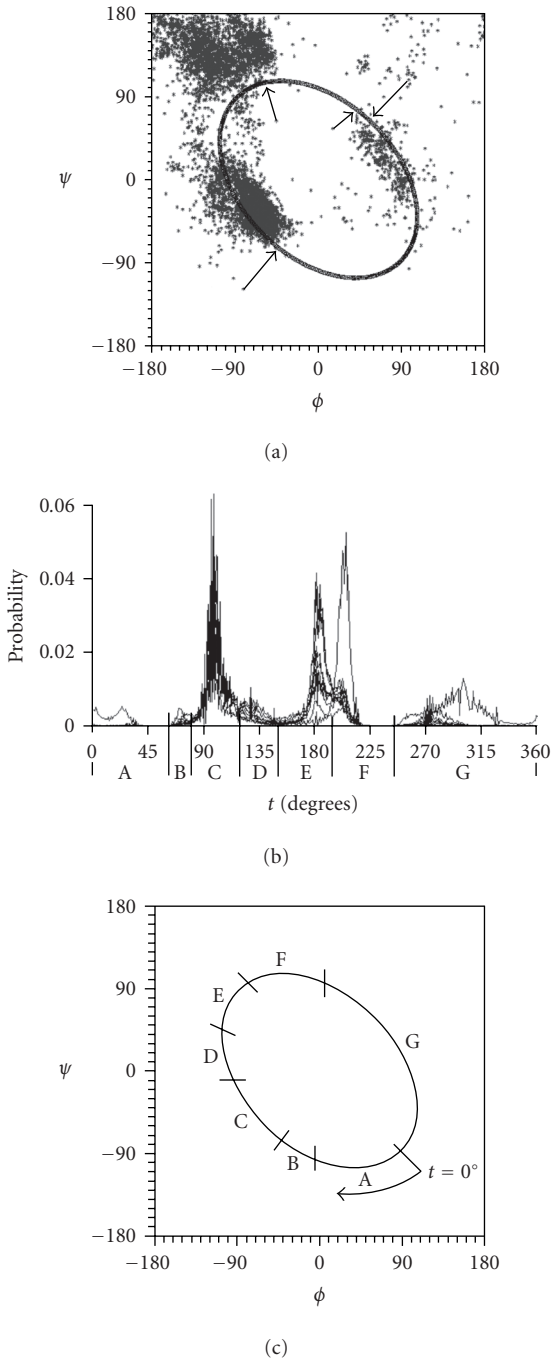


FIGURE 2. Letter codes for structure classification. (a) The ellipse-path-limited conformational subspace in relation to ϕ, ψ angles as they appear in real proteins. Arrows denote the shortest-distance criterion for definition of ϕ, ψ angles belonging to the ellipse for arbitrary selected points. (b) Probability maxima as they appear along the ellipse (starting t-point shown in (c)) and corresponding letter codes for structure identification. (c) Limited conformational subspace with fragments distinguished according to probability maxima shown in (b).

proved that the ellipse-path-limited conformational subspace helped define the initial structure for the energy-minimization procedure, leading to proper, native-like structures without any forms inconsistent with protein-like ones. When the energy-minimization procedure is not sufficient to deliver the proper native-like structure of the protein (which can be seen in Figures 1a and 1j), the additional procedure is necessary (Figure 1i). It is under study now and will be published in the close future.

MATERIAL AND METHODS

Early-stage folding structure classification

All proteins present in PDB (release January 2003) were taken for analysis [54]. Letter codes have been used for sequence identification. A letter code system is introduced in this paper for structure representation in protein early-stage folding (*in silico*) based on the probability distribution of ϕ, ψ angles along the ellipse-path-limited conformational subspace (see "appendix B"). To easily distinguish the structure codes versus sequence codes, the former are printed in bold and the latter in italics in this work.

Comparison of distributions between three-state secondary structures indicated four-amino-acid fragments as the most common ones for α-helices, β-strands, and loops [21, 55]. The tetrapeptide was adopted as the unit for investigation of the sequence-structure relation.

The probability distribution along the ellipse, which is assumed to represent the limited conformational subspace, is the basis for the structure classification introduced in this paper. The profile of the probability distribution (of all amino acids) along the ellipse path is shown in Figure 2. Figure 2a shows the usual distribution of ϕ, ψ angles as found in proteins together with the ellipse path.

The procedure of moving particular ϕ, ψ angles to the ellipse path is also shown in Figure 2a. The shortest distance between particular ϕ, ψ angles (point on the Ramachandran map) and a point belonging to the ellipse path located the ϕ_e, ψ_e (e denotes ellipse belonging) dihedral angles determining the early stage for a particular amino acid of the polypeptide chain. After moving all ϕ, ψ angles to the ellipse path, the profile of the probability distribution can be obtained, as shown in Figure 2b. The t-parameter is the ellipse parameter present in the equation shown in "appendix A." The t-parameter equal to zero represents the point $\phi = 90°$ and $\psi = -90°$ on the Ramachandran map and increases clockwise, as is shown in Figure 2c. Seven probability maxima can be distinguished in this profile. Each of them is letter coded.

This coding system was applied to classify the structures of all proteins analyzed. The codes introduced according to the probability distribution shown in Figure 2b are interpreted as follows: *C* (t-value range) represents right-handed helical structures, *E* represents β-structural forms, and *G* represents left-handed helices. The β-structural forms are differentiated (some amino acids like Ala, Ser, Asp reveal two probability maxima [50]); this is why code *F* also represents β-like structures.

Although all other letters represent structural forms not identified in the traditional classification, the presence of probability maxima suggests the need to distinguish these categories (code A mostly for Pro and Gly, code B represented mostly by Asn and Asp, and code D characteristic for Tyr and Asn, to take a few examples).

The contingency table

A window size of four amino acids (analogous to the open reading frame in nucleotide identification) with one amino acid step (overlapping system) was applied to code the sequences and structures in proteins. Potentially 160 000 (20^4) different sequences for tetrapeptides can occur (columns). Taking seven different structural forms for each amino acid in a tetrapeptide, 2401 (7^4) structural forms can be distinguished for a tetrapeptide (rows). These numbers give an idea of the size of the contingency table under consideration. For all cells, probability values of p^t, p^c, and p^r were calculated as follows:

$$p_{ij}^t = \frac{n_{ij}}{N^t}, \tag{1}$$

$$p_{ij}^c = \frac{n_{ij}}{N_j^c}, \tag{2}$$

$$p_{ij}^r = \frac{n_{ij}}{N_i^r}, \tag{3}$$

where i denotes a particular structure (row), j denotes a particular sequence (column), n_{ij} is the number of polypeptide chains belonging to the ith structure and representing the jth sequence, N^t is the total number of ORFs, and N_j^c and N_i^r denote the number of ORFs belonging to a particular ith structure and jth sequence, respectively. The table expressing all probabilities (p_{ij}^t, p_{ij}^c, and p_{ij}^r) is available on request at http://www.bioinformatics.cm-uj.krakow.pl/earlystage/. All values are expressed on a logarithmic scale because of the very low probability values in the cells of the table.

Information entropy as a measure of sequence-to-structure and structure-to-sequence predictability

High values of probability calculated as above (relative to potential probability values) can disclose highly coupled pairs of structure and sequence. Ranking the probability values can extract the highly determined relations for both sequence-to-structure and structure-to-sequence.

Structural predictability can also be measured using informational entropy calculation. According to Shannon's definition [56], the amount of information can be calculated as follows:

$$I_i = N \log_2 p_i, \tag{4}$$

where I_i expresses the amount of information (in bits) dependent on p_i—the probability of event i. This definition is very useful for measuring the amount of information carried by a particular simple (elementary) event. In the case of a complex event, for which few solutions are possible, informational entropy can be calculated, expressing the level of uncertainty in predicting the solution. Informational entropy according to Shannon's definition is as follows:

$$\mathrm{SE} = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{5}$$

where n is the number of possible solutions for a particular event. N denotes the number of possible solutions for the event under consideration (number of elementary events).

SE reaches its maximum value for all p_i equal to each other, that is, each ith solution is equally probable for the event under consideration and no solution is preferred. The maximum value depends on the number of possible solutions for the event (n).

SE equal to zero (or 1.0) represents the determinate case in which only one solution is possible. The higher the difference between $\overset{\max}{\mathrm{SE}}$ and SE, the higher the degree of determinability in the given case. A high $\overset{\max}{\mathrm{SE}} - \mathrm{SE}$ value means that the case is realized by a few solutions and that some of them occur with higher probability, which can be interpreted as a case with higher determinability (biased event).

SE, $\overset{\max}{\mathrm{SE}}$, and the values of the differences between them can be calculated for all rows SE^r (structural preferences versus amino acid sequence) and for columns SE^c (sequence preference for a particular structural form) in the contingency table. SE^r allowed extraction of structures highly determined by the sequence; SE^c extracted structures highly attributed to a particular sequence.

The SE calculation performed for each column (particular sequence) in the contingency table was calculated as follows:

$$\mathrm{SE}_j^c = -\sum_{i=1}^{N_j^0} p_{ij}^c \log_2 p_{ij}^c, \tag{6}$$

where SE_j^c denotes informational entropy for the j-column, i denotes a particular row (structure), N_j^0 is the number of nonzero cells in the j-column, and p_{ij}^c is calculated according to (2).

The value SE_j^c as calculated according to (6) measures the level of uncertainty in predicting structure for the jth sequence. The closer the SE value to zero, the higher the degree of chance in prediction.

$\overset{\max}{\mathrm{SE}}$ expresses quantitatively the level of uncertainty in the most difficult case for making a decision. For the j-column (sequence):

$$\overset{\max}{\mathrm{SE}}_j^c = -\sum_{i=1}^{N_j^0} \overset{\max}{p_{ij}^c} \log_2 \overset{\max}{p_{ij}^c}, \tag{7}$$

Table 1. Scheme of the sequence-structure contingency table. Symbols explained in text.

Structure	Sequence					
	1	2	⋯	j	⋯	160 000
1	$n_{11}, p^t_{11}, p^c_{11}, p^r_{11}$	$n_{12}, p^t_{12}, p^c_{12}, p^r_{12}$	⋯	$n_{1j}, p^t_{1j}, p^c_{1j}, p^r_{1j}$	⋯	N^r_1
2	$n_{21}, p^t_{21}, p^c_{21}, p^r_{21}$	$n_{22}, p^t_{22}, p^c_{22}, p^r_{22}$	⋯	$n_{2j}, p^t_{2j}, p^c_{2j}, p^r_{2j}$	⋯	N^r_2
⋯	⋯	⋯	⋯	⋯	⋯	⋯
i	$n_{i1}, p^t_{i1}, p^c_{i1}, p^r_{i1}$	$n_{i2}, p^t_{i2}, p^c_{i2}, p^r_{i2}$	⋯	$n_{ij}, p^t_{ij}, p^c_{ij}, p^r_{ij}$	⋯	N^r_i
⋯	⋯	⋯	⋯	⋯	⋯	⋯
2 401	N^c_1	N^c_2	⋯	N^c_j	⋯	N^t

where $\overset{\max}{\mathrm{SE}}{}^c_j$ denotes maximum informational entropy for the j-column, i denotes a particular row (structure), N^0_j is the number of nonzero cells in the j-column, and $\overset{\max}{p}{}^c_{ij}$ denotes the value of probability in a column under the assumption that all nonzero cells are equally represented (the principal condition for $\overset{\max}{\mathrm{SE}}$). In other words, for all nonzero cells ($i = 1, \ldots, N^0_j$) in the j-column $\overset{\max}{p}{}^c_{ij}$ can be calculated as follows:

$$\overset{\max}{p}{}^c_{ij} = \frac{1}{N^0_j}. \tag{8}$$

Thus the difference between two quantities ((6) and (7)) can be used as the "distance" between the most difficult situation (all solutions equally possible—random solution) and the situation observed in the case under consideration. For the j column

$$\Delta\,\mathrm{SE}^c_j = \overset{\max}{\mathrm{SE}}{}^c_j - \mathrm{SE}^c_j. \tag{9}$$

Analogous calculations for rows (sequences) were performed. For each i-row, the value of SE^r_i, $\overset{\max}{\mathrm{SE}}{}^r_i$, and $\Delta\,\mathrm{SE}^r_i$ was calculated.

RESULTS

Structures coded according to the introduced system

Structures of all proteins present in the PDB (release January 2003) [56] were analyzed. The ϕ, ψ angles were calculated for each amino acid. The ϕ_e, ψ_e angles were calculated according to the shortest distance versus the ellipse. A letter code was assigned for each amino acid according to the ellipse path fragment. Since the tetrapeptide was used as the structural unit, four letters coded one structural unit. The overlapping reading frame system was applied, which means that one amino acid step was applied in structure classification. The maximum combination of seven letter codes for a four-letter string is equal to 2401. This means that 2401 different four-letter strings were expected to be found. It turned out that only 2397 different strings were found in real proteins. Since there are 20 amino acids and four amino acids were taken for the unit, 160 000 different sequences of tetrapeptides were expected; 146 940 different sequences were found in the proteins under consideration.

Contingency table

Each tetrapeptide found in proteins was described by a four-letter string expressing the sequence and a four-letter string expressing the structure. Each tetrapetide with a known sequence and known structure can be ordered in the form of a table. The rows of the table represent structures and the columns represent sequences. Finally a $2397 \times 146\,940$ table was constructed. To distinguish the structure codes from sequence codes, sequence codes are in bold capital letters and structure codes in italics. The scheme of the contingency table is presented in Table 1. The total number of tetrapeptides in the analyzed database was found to be 1 529 987. Global analysis of the contingency table shows that the maximum number of different structures attributed to the same tetrapeptide is 144. This tetrapeptide appeared to be of the sequence *GSAA*. The maximum number of different sequences was found for α-helix (*CCCC*: 90 587) and for β-structure (*EEEE*: 47 809). Four structures were not found in the library: *ABAB*, *ABBD*, *ABFB*, *DBAB*.

Information entropy calculation

SE, $\overset{\max}{\mathrm{SE}}$, and the value of the difference between these two quantities (ΔSE) were calculated according to the procedure presented in "material and methods." They can be calculated for columns (sequences) and for rows (structures) separately. The calculation of SE^c_j for the j-column expresses the information entropy related to the structural differentiation of a particular sequence. The calculation of SE^r_i for the i-row in the contingency table expresses the sequential differentiation for a particular structure. $\overset{\max}{\mathrm{SE}}$ according to information entropy characteristics expresses the entropy for the case in which each of all the nonzero cells represents equal probability. For $\mathrm{SE}^c_j = \overset{\max}{\mathrm{SE}}{}^c_j$, all structures for a particular sequence are equally probable. Equal probability for a set of elementary events (different structures) represents the random situation. The bigger the difference $\overset{\max}{\mathrm{SE}} - \mathrm{SE}$, the more deterministic the case. This is why the difference (Δ SE) between SE and $\overset{\max}{\mathrm{SE}}$ was taken to measure the degree of structure-to-sequence (or vice versa) determination.

The interpretation of Tables 2 and 3 is as follows. The structural predictability for a particular sequence can be

Table 2. Sequence-to-structure relation measured according to the value of the difference (ΔSE^c) between entropy of information (SE^c) calculated for the probability values found in the contingency table (particular column) and maximum entropy of information ($\overset{max}{SE^c}$), which (according to the characteristics of entropy of information) is reached for equal probability values in each nonzero cell in a particular column.

Sequence	Structure	SE^c (bit)	$\overset{max}{SE^c}$ (bit)	ΔSE^c (bit)
AAAA	*CCCC*	2.29	6.44	4.15
GDSG	*GCFG*	1.57	5.49	3.92
AVRR	*CCCC*	1.04	4.95	3.91
LAAA	*CCCC*	1.77	5.61	3.84
EAEL	*CCCC*	1.37	5.21	3.83
LDKA	*CCCC*	1.30	5.09	3.78
DAAV	*CCCC*	0.69	4.46	3.77
AKLK	*CCCC*	0.76	4.52	3.77
DSGG	*CFGF*	1.97	5.73	3.76
ELAA	*CCCC*	1.30	5.04	3.75

Table 3. Structure-to-sequence relation measured according to the value of the difference (ΔSE^r) between entropy of information (SE^r) calculated for the probability values found in the contingency table (particular column) and maximum entropy of information ($\overset{max}{SE^r}$), which (according to the characteristics of entropy of information) is reached for equal probability values in each nonzero cell in a particular column.

Structure	Sequence	SE^r (bit)	$\overset{max}{SE^r}$ (bit)	ΔSE^r (bit)
GCFG	**GDSG**	4.82	7.99	3.17
AEED	**GLRL**	3.86	6.81	2.95
BACE	**GGAE**	2.20	5.09	2.89
EAEG	**IGIG**	4.79	7.68	2.89
AEGE	**GIGH**	4.74	7.63	2.89
BFBE	**PEPV**	2.28	5.13	2.85
AEGD	**GNES**	2.09	4.91	2.82
EBCB	**ELPD**	3.68	6.38	2.70
EBFB	**FBEP**	2.57	5.17	2.60
AFFP	**GFRN**	2.03	4.58	2.55

estimated in the first case, and the predictability of the sequence for a particular structure in the latter case. The results for only the top ten structures and top ten sequences are shown in Tables 2 and 3.

Its highest structural predictability for a particular sequence confirms polyalanine as a highly probable helical structure. Generally, the highly predictable structures for particular sequences are helical forms (Table 2).

The sequence predictability for particular structural forms displayed a quite unexpected regularity. The structures representing irregular structural forms appeared to reveal the strongest entropy decrease versus the random distribution of sequences. This can be seen analyzing the letter codes for the structures (Table 3).

The top ten structures presented in Table 3 are also shown in Figure 3. In summary, one can say that when a particular irregular structural form is expected in a protein, there are preferable sequences to build these irregular motifs; they are shown in Table 3. This seems to be of particular relevance for threading procedures oriented to the production of new proteins not observed in nature.

DISCUSSION

Particular classes of amino acid relations to particular structural forms in proteins were recently found to solve the problem of structure predictability [57]. All papers concerning this subject linked sequence with structure as it appears in the final native form of the protein. The model introduced in this paper represents an approach to the relation between sequence and structure in the early-stage folding structural form; the bases for the model are presented in detail elsewhere [48, 49, 50], and verified by BPTI [51], ribonuclease [50], hemoglobin [52], and lysozyme [53] folding. The (*in silico*) early-stage structures of these proteins can be found in the corresponding publications.

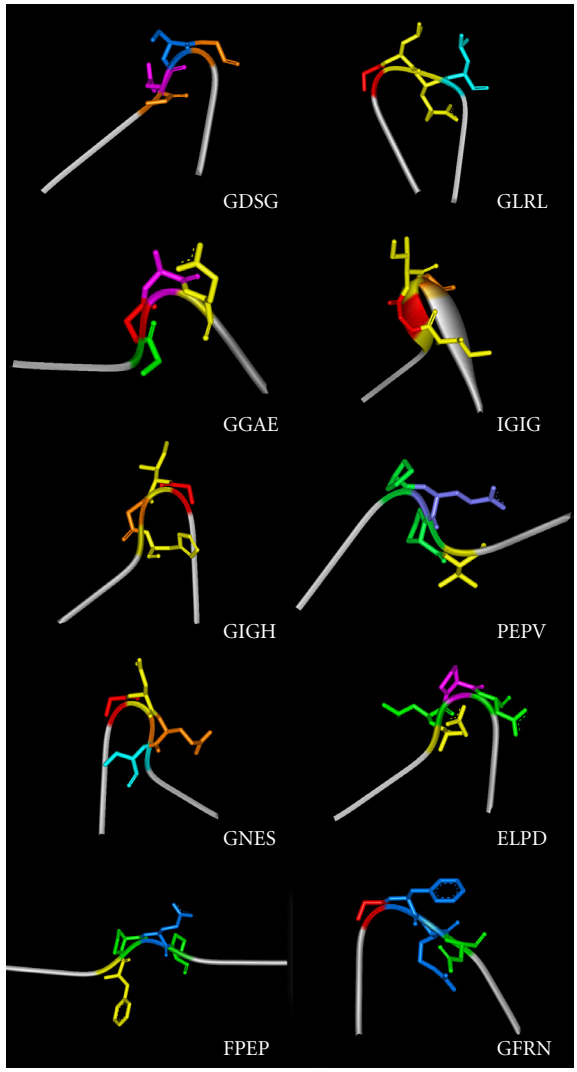


FIGURE 3. Structures of tetrapeptides with highest structure-to-sequence determinability as found using informational entropy calculation (see "material and methods" and Table 1). Gray terminal fragments represent the extended form of polyalanine (tetrapeptides) to emphasize the mutual spatial orientation of terminal fragments. Other colors distinguish ellipse fragments as follows: red (*A*), green (*B*), violet (*C*), sky-blue (*D*), yellow (*E*), dark blue (*F*), orange (*G*). The data for creation of these structures is given in Table 1 and Figure 2.

Several algorithms for quantitatively assigning α-helix, β-strand, and loop regions for proteins with known structure have been developed [58, 59, 60, 61]. The three-dimensional model presented in this paper shows that it is enough to select seven fragments of the ellipse with well-defined probability maxima to be able to predict the early-stage structural form.

The high structure-to-sequence relation found for loops (Table 3, Figure 3) may be particularly important, since a recent survey of 31 genomes indicated that disordered segments longer than 50 residues are very prevalent [62]. Helices, sheets, and turns together account for only about 50%–55% of all protein structure on average [63]; the remaining structures are classified as several types of loops [63, 64]. Current estimations suggest that over 50% of proteins in eukaryotes may carry unconstructed regions of more than 40 residues in length [65], while less than 1% of the proteins in the PDB contains such long disordered regions. These observations taken together imply that many proteins with disordered regions would be unlikely to form crystals [66]. Proteins containing long, disordered segments under physiological conditions are frequently involved in regulatory functions [67], and the structural disorder may be relieved upon binding of the protein to its target molecule [68, 69]. Intrinsically unconstructed proteins and regions, which are also known as natively unfolded and intrinsically disordered, differ from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, hydrophobicity, charge, flexibility, and type and rate of amino acid substitutions over evolutionary time [66]. Compared to highly ordered secondary structure regions, the loops and turns are more difficult to identify due to the absence of hydrogen bonding and repeating backbone dihedral angle patterns [70]. The first computational tool indicating the predictability of disordered regions from protein sequence [71] was a neural network predictor (PONDR). Several other disorder predictors have been published since then [72, 73, 74]. Statistically based turn propensity used over a four-residue window was described [75]. The inverse folding problem is the design of protein sequences that have a desired structure [76, 77]. It is impossible to mention even a small part of the papers dealing with the sequence-to-structure relation. Recently, it was concluded that the probability of any state (ϕ,ψ) is influenced by the full sequence and not only by the local structure [78].

A genome-scale fold recognition program exploring the knowledge-based structure-derived score function for a particular residue was proposed incorporating three terms: backbone torsion, buried surface, and contact energy [79].

Unlike many others, our model, dual in nature, incorporating sequential and structural information, predicts sequence-to-structure as well as structure-to-sequence.

The contingency table was independently analyzed using another statistics-related method (Meus J, Stefaniak J. The Z coefficient as a measure of dependence in contingency tables (unpublished data), Meus J, Brylinski M, Piwowar P, et al. A tabular approach to the sequence-to-structure relation in proteins (unpublished data)). High accordance was found between the results presented in this paper and in the statistical analysis: the top ten sequences and structures presented in Table 1 were found to be among the most highly correlated, both in sequence-to-structure and in structure-to-sequence, on the ranking list created by the alternate calculation method. The order of the two ranking lists is very similar, additionally confirming the reliability of the model presented.

Aside from early-stage structure prediction, the contingency table presented may contribute to conventional secondary structure prediction, local and supersecondary structure prediction, location of transmembrane regions in proteins, location of genes, or sequence design.

The list of highly determinable tetrapeptides (in sequence-to-structure and structure-to-sequence relations) also allowed the SPI (structure predictability index) scale to be defined [80]. Applied to amino acid sequences, this scale helps to measure the degree of difficulty of structure prediction for a particular amino acid sequence without knowledge of the final, native structure of the protein.

The sequence-to-structure and structure-to-sequence contingency tables, which is created on the basis of all proteins of known structure (step-back procedure), can be used to create the early-stage folding (*in silico*) structure. Applied to other (late-stage folding) procedures, it presumably can enable protein structure prediction. The early-stage form was used as the object for comparison to simplify the presentation of the structure (seven possibilities). The SPI (structure predictability index) parameter, attributed to any amino acid sequence, allows estimation of the degree of difficulty in structure prediction. The probability values (which can be higher or lower) taken from particular cells of the contingency table can tell how offen a particular structure occurs in the protein database so far. The information entropy-based classification presented in this paper allows highly distributed structural forms to be distinguished for a particular tetrapeptide sequence.

APPENDIX A

The main assumption for the model presented below is that all structural forms of polypeptides in proteins can be treated as helical. The β-structure in this approach is a helix with a very large radius of curvature. The radius of curvature depends on the V-angle, which expresses the dihedral angle between two sequential peptide bond planes. The quantitative analysis of the relation between these two parameters (V and R) used the following procedure.

(1) The structure of the alanine pentapeptide was created for each 5° grid point on the Ramachandran map. Each alanine present in the pentapetide represented the ϕ, ψ angles appropriate for a particular grid point.

(2) Before the parameters (R, V) were calculated, all structures (for each grid point) were oriented in a unified way: the averaged position of the carbonyl oxygen atoms and the averaged position of carbonyl carbon atoms determined the Z-axis.

(3) The radius of curvature was calculated for projections of $C\alpha$ atoms on the xy plane. The radius of curvature for extended (and β-structural) forms is very large (theoretically infinite). This is why the natural logarithmic scale was introduced to express the magnitude of R.

(4) The V-angle was calculated as the difference between the tilt of the central peptide bond plane and the tilt of two (averaged) neighboring peptide bond planes.

The Ramachandram map expressing the V-angle distribution and R-radius of curvature (in ln scale) is shown in Figure 6.

The ($\ln R$) dependence on the V-angle for structures representing low-energy conformations is shown in Figure 4. The approximation function found for this relation is as follows:

$$\ln(R) = 3.4 * 10^{-4} * V^2 - 2.009 * 10^{-2} * V + 0.848. \quad \text{(A.1)}$$

The distribution of ϕ, ψ angles of structures that satisfy the above equation is shown in Figure 5. The ellipse path found based on this distribution is as follows:

$$\begin{aligned} \phi &= -A\cos(t) - B\sin(t), \\ \psi &= A\cos(t) - B\sin(t), \end{aligned} \quad \text{(A.2)}$$

where A and B are long and short ellipse diagonals, respectively.

APPENDIX B

The sequence of amino acids in polypeptide determines its structural form. This expression can be understood also as follows. The amount of information carried by an amino acid sequence is comparable to the amount of information necessary to predict its structure.

The amount (bit) of information carried by a particular amino acid can be calculated using Shannon's equation

$$I_i(p_i) = -\log_2 p_i, \quad \text{(B.1)}$$

where p_i expresses the probability of the ith amino acid's presence in a sequence.

Assuming all amino acids occur with the same probability (1/20), the amount of information can be calculated.

The amount of information necessary to predict a particular structure (expressed by ϕ_i, ψ_i dihedral angles) for the ith amino acid can also be calculated as follows (using the same Shannon's equation):

$$I_i^{\phi\psi} = -\log_2 p_i^{\phi\psi}, \quad \text{(B.2)}$$

where $p_i^{\phi\psi}$ expresses the probability of the ith amino acid to represent the ϕ, ψ dihedral angles. Assuming 1° as the step for exploring the Ramachandran map and assuming that the Ramachandran map is flat (all ϕ, ψ angles equally possible), the amount of information I is calculated for $p_i^{\phi\psi}$ equal to $1/(359 * 359)$.

This simple comparison shows that the big difference makes the situation highly nonequilibrated.

The value of p_i is different from 1/20 in real proteins because the frequency of amino acids differs.

The value of $p_i^{\phi\psi}$ also depends on the amino acid under consideration. The assumption of equal probability of

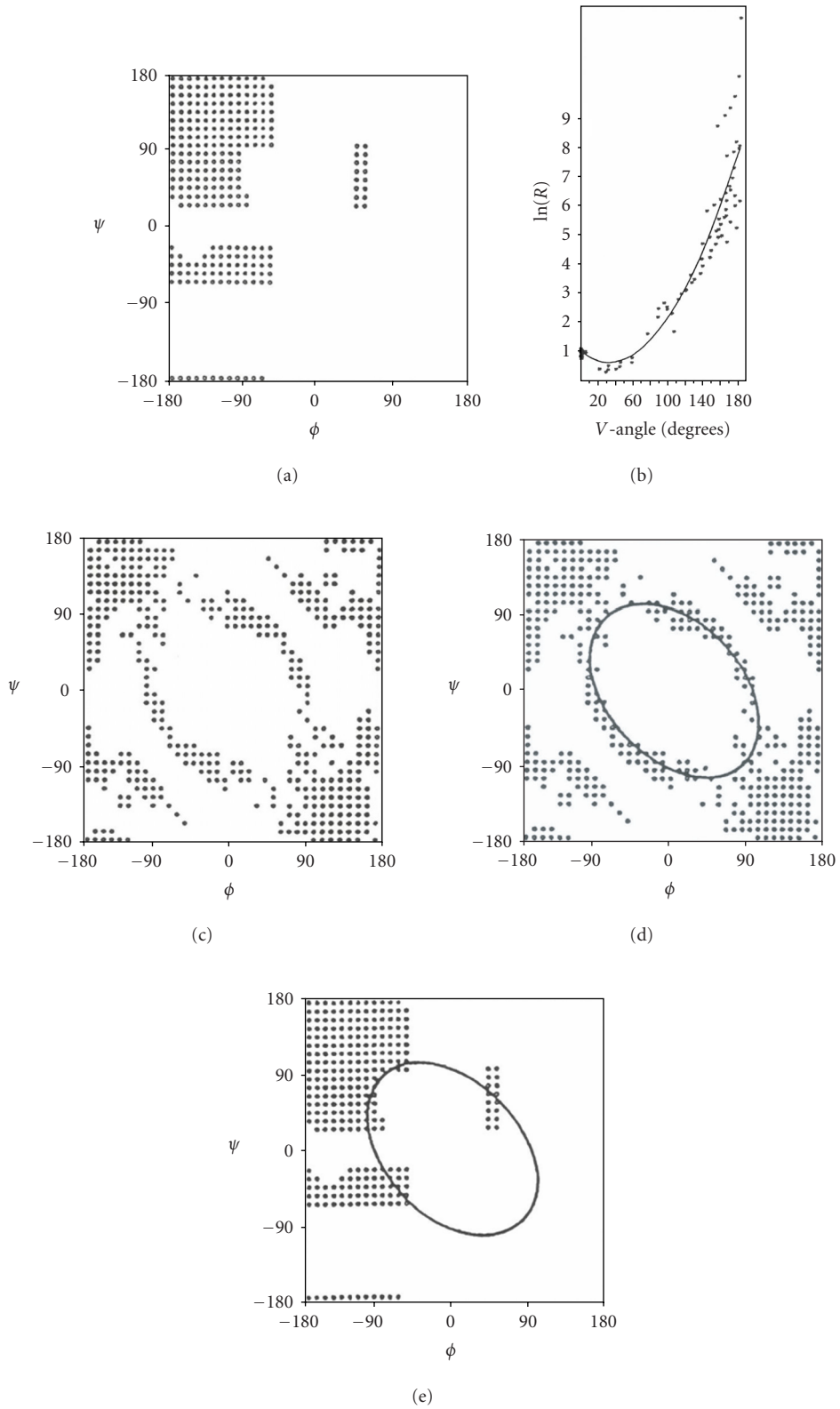
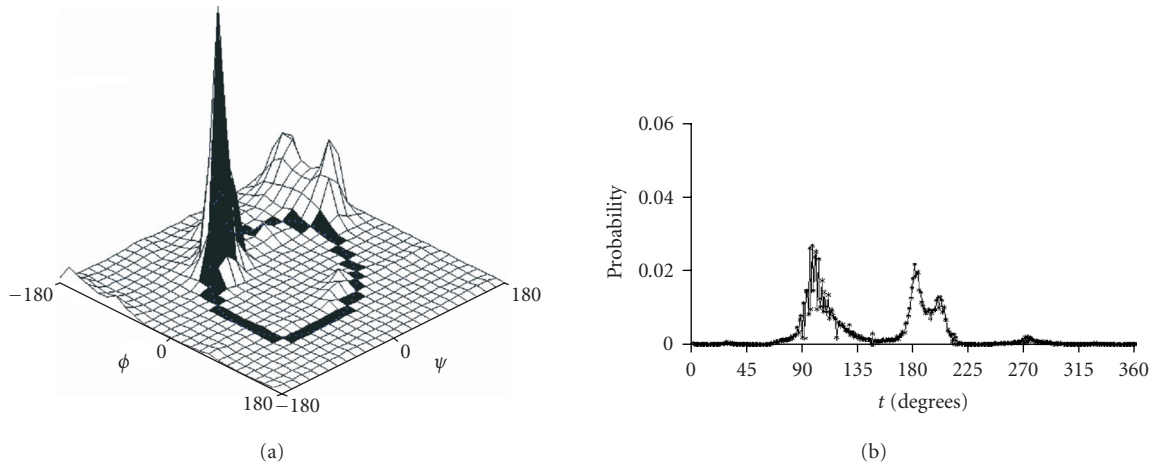
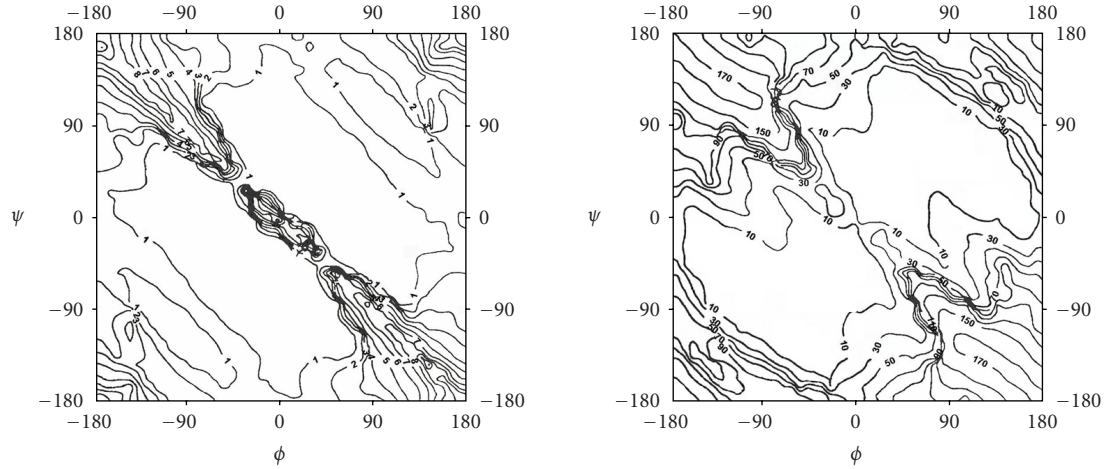


FIGURE 4. Ellipse path determination. (a) ϕ, ψ map with low-energy area distinguished, (b) $\ln(R)$ as a function of V-angle for grid points shown in (a), (c) ϕ, ψ map with grid points, where the structure satisfies (1), (d) proposed ellipse path, (e) low-energy areas linked by ellipse.



(a) (b)

FIGURE 5. ϕ, ψ angles distribution of serine: (a) all over the Ramachandran map, black line distinguishes the ellipse path, (b) after moving all ϕ, ψ angles toward the ellipse path. The variable called t expresses the variable in the ellipse equation (A.2). Zero value of t represents the point $\phi = 90^\circ$, $\psi = -90^\circ$ and then increases clockwise along the ellipse. The probability profiles for each amino acid representing the ϕ, ψ angles in real proteins after transforming them to the ellipse-path-limited conformational subspace (shortest distance criterion) are presented previously [50].



(a) (b)

FIGURE 6. Distribution of geometrical parameters all over the ϕ, ψ map. (a) Radius of curvature "R" on natural logarithmic scale. (b) Dihedral angle "V" between two sequential peptide bond planes.

ϕ, ψ angles cannot be accepted. Predicting particular ϕ, ψ angles is relatively easy for proline and most difficult for glycine. Prediction of particular ϕ, ψ angles is connected with the selection decision. This means selection of ϕ, ψ from among $359 * 359$ possible solutions. Moreover, particular ϕ, ψ angles are not equally possible. With information entropy measuring the degree of uncertainty in ϕ, ψ angles, selection (according to Shannon's equation) is calculated as follows:

$$\mathrm{SE}_i = -\sum_{i=1}^{359*359} p_i \log_2 p_i, \tag{B.3}$$

where index i denotes the amino acid under consideration, p_i denotes the probability of occurrence of particular ϕ, ψ angles calculated for the ith amino acid, N denotes the number of grid points (depending on the step size for ϕ, ψ angles all over the Ramachandran map), and SE_i expresses the mean value (quantity) of information (bit) necessary to select one solution from among the number that represents the complete event space ($359 * 359$ in our case). The mean value takes into account the different probabilities for different ϕ, ψ angles and also the dependence on the amino acid under consideration (ith). SE can be interpreted as a scale to measure the predictability

Table 4. Amount of information (I_i (bit)) carried by a particular amino acid, calculated on the basis of the frequency and amount of information ($SE_i^{\phi_e\psi_e}$ (bit), ϕ_e, ψ_e denote ϕ, ψ angles belonging to the ellipse) necessary to predict the structure belonging to the ellipse path (early-stage folding conformational subspace) with 10° step of t-angle precision (see ellipse equation in "appendix A"). Detailed analysis of the data shown in this table can be found elsewhere [50].

Amino acid	Amount of information carried by amino acid	Averaged amount of information necessary to predict the ellipse-belonging structure
	I_i (bit)	$SE_i^{\phi_e\psi_e}$ (bit)
Gly	3.805	7.806
Asp	4.117	7.073
Leu	3.492	6.438
Lys	3.908	6.789
Ala	3.662	6.409
Ser	4.095	6.975
Asn	4.545	7.267
Glu	3.833	6.520
Thr	4.196	6.720
Arg	4.249	6.677
Val	3.886	6.233
Gln	4.663	6.676
Ile	4.151	6.208
Phe	4.713	6.617
Tyr	4.941	6.685
Pro	4.442	6.124
His	5.477	6.965
Cys	5.544	6.937
Met	5.614	6.494
Trp	6.236	6.581

characteristic for a particular amino acid. It was shown that the SE scale places Gly and Pro at opposite positions on the ranking (scoring) list of amino acids. The $10 * 10$ step for ϕ, ψ angles precision prediction still needs a large amount of information to be equilibrated with the amount of information carried by a particular amino acid (in this case N is equal to $35 * 35$).

Analysis of the ellipse path from the point of view of SE calculation reveals that this limited conformational subspace (with 10° steps along the ellipse expressed as N as in (B.4)) satisfies the condition of balancing (Table 4) the amount of information carried by amino acid and the amount of information necessary for selection of the structure belonging to the ellipse path representing the limited conformational subspace with 10° precision.

$$SE_i = -\sum_{i=1}^{360/N} p_i \log_2 p_i, \tag{B.4}$$

where p_i denotes the probability value for a particular point on the ellipse (particular t-parameter), and N denotes the number of points selected (it is coupled with the t-parameter step size).

The ellipse path presented in "appendix A" appeared to satisfy two important conditions. (i) Almost all structurally important forms of polypeptide are present in this conformational subspace; and (ii) the amount of information carried by the amino acid and the amount of information needed to predict a particular structural form belonging to the conformational subspace are equilibrated. Details on the information problem can be found elsewhere [50]. Figure 5a shows the relation between the ϕ, ψ angles of Ser distribution all over the Ramachandran map, with the ellipse path distinguished by a black line. The distribution of the ϕ, ψ angles of Ser after moving them toward the ellipse path is shown in Figure 5b. The overlapping of the probability profiles of all amino acids is shown in Figure 1b.

ACKNOWLEDGMENTS

We wish to thank Professor Marek Pawlikowski (Faculty of Chemistry, Jagiellonian University) for fruitful discussions. The complete contingency table may be obtained by contacting the authors. This work was financially supported by Collegium Medicum Grants 501/P/133/L, WŁ/222/P/L.

REFERENCES

[1] Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA.* 1984;81(4):1075–1078.

[2] Maxfield FR, Scheraga HA. Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry.* 1979;18(4):697–704.

[3] Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol.* 1987;195(4):957–961.

[4] Benner SA. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv Enzyme Regul.* 1989;28:219–236.

[5] Shortle D. Prediction of protein structure. *Curr Biol.* 2000;10(2):49–51.

[6] Efimov AV. Role of connections in the formation of protein structures, containing 4-helical segments. *Mol Biol (Mosk).* 1982;16(2):271–281.

[7] Efimov AV. A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. *FEBS Lett.* 1984;166(1):33–38.

[8] Lim VI. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol.* 1974;88(4):873–894.

[9] Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry.* 1974;13(2):211–222.
[10] Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry.* 1974;13(2):222–245.
[11] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol.* 1978;120(1):97–120.
[12] Garnier J, Robson, B. The GOR method for predicting secondary structures in proteins. In: Fasman GD, ed. *Prediction of Protein Structure and the Principles of Protein Conformation.* New York, NY: Plenum Press; 1989:417–465.
[13] Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. Secondary structure prediction: combination of three different methods. *Protein Eng.* 1988;2(3):185–191.
[14] Levin JM, Robson B, Garnier J. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* 1986;205(2):303–308.
[15] Nishikawa K, Ooi T. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim Biophys Acta.* 1986;871(1):45–54.
[16] Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. *J Mol Biol.* 1993;232(4):1117–1129.
[17] Holley LH, Karplus M. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA.* 1989;86(1):152–156.
[18] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999;292(2):195–202.
[19] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol.* 1988;202(4):865–884.
[20] Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden Markov model. *Comput Appl Biosci.* 1993;9(2):141–146.
[21] Bystroff C, Thorsson V, Baker D. A hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol.* 2000;301(1):173–190.
[22] Bystroff C, Shao Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics.* 2002;18(1):54–61.
[23] Stultz CM, White JV, Smith TF. Structural analysis based on state-space modeling. *Protein Sci.* 1993;2(3):305–314.
[24] Aurora R, Srinivasan R, Rose GD. Rules for alpha-helix termination by glycine. *Science.* 1994;264(5162):1126–1130.
[25] Harper ET, Rose GD. Helix stop signals in proteins and peptides: the capping box. *Biochemistry.* 1993;32(30):7605–7609.
[26] Presnell SR, Cohen BI, Cohen FE. A segment-based approach to protein secondary structure prediction. *Biochemistry.* 1992;31(4):983–993.
[27] Zhou HX, Lyu P, Wemmer DE, Kallenbach NR. Alpha helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: evidence for an N terminal "capping box". *Proteins.* 1994;18(1):1–7.
[28] Bonneau R, Strauss CE, Rohl CA, et al. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol.* 2002;322(1):65–78.
[29] Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol.* 1998;281(3):565–577.
[30] Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA.* 1996;93(12):5814–5818.
[31] Crawford IP, Niermann T, Kirschner K. Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins.* 1987;2(2):118–129.
[32] Russell RB, Breed J, Barton GJ. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Lett.* 1992;304(1):15–20.
[33] Benner SA, Cohen MA, Gerloff D. Predicted secondary structure for the Src homology 3 domain. *J Mol Biol.* 1993;229(2):295–305.
[34] Levin JM, Pascarella S, Argos P, Garnier J. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* 1993;6(8):849–854.
[35] Hansen JE, Lund O, Nielsen JO, Brunak S, Hansen JE. Prediction of the secondary structure of HIV-1 gp120. *Proteins.* 1996;25(1):1–11.
[36] Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. *J Mol Biol.* 1997;268(1):31–36.
[37] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol.* 1993;232(2):584–599.
[38] Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol.* 1994;235(1):13–26.
[39] Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins.* 2003;53(6):436–456.
[40] Liwo A, Czaplewski C, Pillardy J, Scheraga, HA. Cummulation-based expression for the multibody terms for the correction between local and electrostatic interaction in the united residue force field. *J Chem Phys.* 2001;115:2323–2347.
[41] Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pillardy J, Scheraga HA. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application

to the UNRES force field. *Proc Natl Acad Sci USA.* 2002;99(4):1937–1942.
[42] Mezei M. A novel fingerprint for the characterization of protein folds. *Protein Eng.* 2003;16(10):713–715.
[43] Fernandez A, Colubri A, Appignanesi G, Burastero T. Coarse semiempirical solution to the protein folding problem. *Physica A.* 2001;293:358–384.
[44] Sosnick TR, Berry RS, Colubri A, Fernandez A. Distinguishing foldable proteins from nonfolders: when and how do they differ? *Proteins.* 2002;49(1):15–23.
[45] Pappu RV, Srinivasan R, Rose GD. The flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci USA.* 2000;97(23):12565–12570.
[46] Colubri A. Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report. *J Biomol Struct Dyn.* 2004;21(5):625–638.
[47] Alonso DO, Daggett V. Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Sci.* 1998;7(4):860–874.
[48] Roterman I. Modelling the optimal simulation path in the peptide chain folding—studies based on geometry of alanine heptapeptide. *J Theor Biol.* 1995;177(3):283–288.
[49] Roterman I. The geometrical analysis of peptide backbone structure and its local deformations. *Biochimie.* 1995;77(3):204–216.
[50] Jurkowski W, Brylinski M, Konieczny L, Wiiniowski Z, Roterman I. Conformational subspace in simulation of early-stage protein folding. *Proteins.* 2004;55(1):115–127.
[51] Brylinski M, Jurkowski W, Konieczny L, Roterman I. Limited conformational space for early-stage protein folding simulation. *Bioinformatics.* 2004;20(2):199–205.
[52] Brylinski M, Jurkowski W, Konieczny L, Roterman I. Limitation of conformational space for proteins—early-stage folding simulation of human α and β hemoglobin chains. *TASK-Quarterly.* 2004;8:413–422.
[53] Jurkowski W, Brylinski M, Konieczny L, Roterman I. Lysozyme folded in silico according to the limited conformational sub-space. *J Biomol Struct Dyn.* 2004;22(2):149–158.
[54] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res.* 2000;28(1):235–242.
[55] Zhu ZY, Blundell TL. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J Mol Biol.* 1996;260(2):261–276.
[56] Shannon CEA. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27:379–423.
[57] Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. *Proteins.* 2002;48(3):463–486.
[58] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 1983;22(12):2577–2637.
[59] Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins.* 1988;3(2):71–84.
[60] Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. *J Mol Biol.* 1977;114(2):181–239.
[61] Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins.* 1989;6(1):46–60.
[62] Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model.* 2001;19(1):26–59.
[63] Leszczynski JF, Rose GD. Loops in globular proteins: a novel category of secondary structure. *Science.* 1986;234(4778):849–855.
[64] Ring CS, Kneller DG, Langridge R, Cohen FE. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol.* 1992;224(3):685–699.
[65] Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins.* 2003;52(4):573–584.
[66] Iakoucheva LM, Dunker AK. Order, disorder, and flexibility: prediction from protein sequence. *Structure (Camb).* 2003;11(11):1316–1317.
[67] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry.* 2002;41(21):6573–6582.
[68] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol.* 2002;12(1):54–60.
[69] Dyson HJ, Wright PE. Intrinsically unconstructed proteins: re-assessing the protein structure-function paradigm. *J Mol Biol.* 1999;293:321–331.
[70] Fetrow JS, Palumbo MJ, Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins.* 1997;27(2):249–271.
[71] Romero P, Obradovic Z, Kissinger CR, et al. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput.* 1998;3:437–448.
[72] Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins.* 2000;41(3):415–427.
[73] Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.* 2003;31(13):3833–3835.
[74] Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 2003;31(13):3701–3708.
[75] Hutchinson EG, Thornton JM. A revised set of potentials for beta-turn formation in proteins. *Protein Science.* 1994;3:2207–2216.
[76] Fischer N, Riechmann L, Winter G. A native-like ar-

tificial protein from antisense DNA. *Protein Eng Des Sel.* 2004;17(1):13–20.

[77] Wie Y, Hecht MH. Enzyme-like proteins from an unselected library of designed amino acid sequences. *Protein Eng Des Sel.* 2004;17:67–75.

[78] Keskin O, Yuret D, Gursoy A, Turkay M, Erman B. Relationships between amino acid sequence and backbone torsion angle preferences. *Proteins.* 2004;55(4):992–998.

[79] Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins.* 2004;55(4):1005–1013.

[80] Brylinski M, Konieczny L, Roterman I. SPI—structure predictability index for proteins. *In Silico Biology.* 2004;5:0022.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 80–86 • DOI: 10.1155/JBB.2005.80

RESEARCH ARTICLE

Functional Clustering Algorithm for High-Dimensional Proteomics Data

Halima Bensmail,[1] Buddana Aruna,[1] O. John Semmes,[2] and Abdelali Haoudi[2]

[1]*Department of Statistic Operation and Management Sciences (SOMS), The University of Tennessee, Knoxville, TN 37996, USA*
[2]*Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23507, USA*

Received 9 September 2004; revised 10 February 2005; accepted 14 February 2005

Clustering proteomics data is a challenging problem for any traditional clustering algorithm. Usually, the number of samples is largely smaller than the number of protein peaks. The use of a clustering algorithm which does not take into consideration the number of features of variables (here the number of peaks) is needed. An innovative hierarchical clustering algorithm may be a good approach. We propose here a new dissimilarity measure for the hierarchical clustering combined with a functional data analysis. We present a specific application of functional data analysis (FDA) to a high-throughput proteomics study. The high performance of the proposed algorithm is compared to two popular dissimilarity measures in the clustering of normal and human T-cell leukemia virus type 1 (HTLV-1)-infected patients samples.

INTRODUCTION

A variety of mass spectrometry-based platforms are currently available for providing information on both protein patterns and protein identity [1, 2]. Specifically, the first widely used such mass spectrometric technique is known as surface-enhanced laser desorption ionization (SELDI) coupled with time-of-flight (TOF) mass spectrometric detection [3, 4, 5]. The SELDI approach is based on the use of an energy-absorbing matrix such as sinapinic acid (SPH), large molecules such as peptides ionize instead of decomposing when subjected to a nitrogen UV laser. Thus, partially purified serum is crystallized with an SPH matrix and placed on a metal slide. Depending upon the range of masses the investigator wishes to study, there are a variety of possible slide surfaces; for example, the strong anion exchange (SAX) or the weak cation exchange (WCX) surface. The peptides are ionized by the pulsed laser beam and then traverse a magnetic-field-containing column. Masses are separated according to their TOFs as the latter are proportional to the square of the mass-to-charge (m/z) ratio. Since nearly all of the resulting ions have unit charge, the mass-to-charge ratio is in most cases a mass. The spectrum (intensity level as a function of mass) is recorded, so the resulting data obtained on each serum sample are a series of intensity levels at each mass value on a common grid of masses (peaks).

Proteomic profiling is a new approach to clinical diagnosis, and many computational challenges still exist. Not only are the platforms themselves still improving, but the methods used to interpret the high-dimensional data are developing as well [6, 7].

A variety of clustering approaches has been applied to high-dimensional genomics and proteomics data [8, 9, 10, 11]. Hierarchical clustering methods give rise to nested partitions, meaning the intersection of a set in the partition at one level of the hierarchy with a set of the partition at a higher level of the hierarchy will always be equal to the set from the lower level or the empty set. The hierarchy can thus be graphically represented by a tree.

Functional data analysis (FDA) is a statistical data analysis represented by smooth curves or continuous functions $\mu_i(t)$, $i = 1, \ldots, n$, where n is the number of observations and t might or might not necessarily denote time but might have a general meaning. Here t denotes the mass (m/z). In practice, the information over $\mu_i(t)$ is collected at a finite number of points, T_i, thus observing the data vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})^t$. The basic statistical model

Correspondence and reprint requests to Abdelali Haoudi, Eastern Virginia Medical School, Department of Microbiology and Molecular Cell Biology, Norfolk, VA 23507, USA, Email: haoudia@evms.edu

of FDA is given by

$$y_{ij} = \hat{\mu}_i(t_{ij}) = \mu_i(t_{ij}) + \epsilon_i(t_{ij}), \quad i = 1, \ldots, n, \ j = 1, \ldots, T_i, \tag{1}$$

where t_{ij} is the mass value at which the jth measurementis taken for the ith function μ_i. The independent disturbance terms $\epsilon_i(t_{ij})$ are responsible for roughness in $\mathbf{y}_i$. FDA has been developed for analyzing functional (or curve) data. In FDA, data consists of functions not of vectors. Samples are taken at time points $t_1, t_2, \ldots$, and regard $\mu_i(t_{ij})$ as multivariate observations. In this sense the original functional y_{ij} can be regarded as the limit of $\mu_i(t_{ij})$ as the sampling interval tends to zero and the dimension of multivariate observations tends to infinity. Ramsay and Silverman [12, 13] have discussed several methods for analyzing functional data, including functional regression analysis, functional principal component analysis (PCA), and functional canonical correlation analysis (CCA). These methodologies look attractive, because one often meets the cases where one wishes to apply regression analysis and PCA to such data. In the following we describe how to use the FDA tools for applying FDA and a new dissimilarity measure to classify the spectra data.

We propose to implement a hierarchical clustering algorithm for proteomics data using FDA. We use functional transformation to smooth and reduce the dimensionality of the spectra and develop a new algorithm for clustering high-dimensional proteomics data.

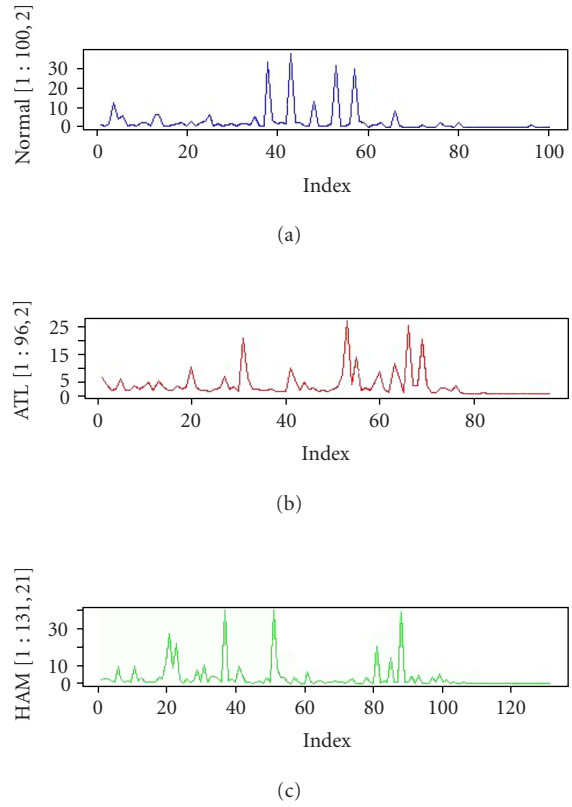


Figure 1. Three cut expressions from a normal, an HAM, and an ATL patient.

MATERIAL AND METHODS

Serum samples from HTLV-1-infected patients

Protein expression profiles generated through SELDI analysis of sera from human t-cell leukemia virus type 1- (HTLV-1)-infected individuals were used to determine the changes in the cell proteome that characterize adult T-cell leukemia (ATL), an aggressive lymphoproliferative disease from HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), a chronic progressive neurodegenerative disease. Both diseases are associated with the infection of T cells by HTLV-1. The HTLV-1 virally encoded oncoprotein Tax has been implicated in the retrovirus-mediated cellular transformation and is believed to contribute to the oncogenic process through induction of genomic instability affecting both DNA repair integrity and cell cycle progression [14, 15]. Serum samples were obtained from the Virginia Prostate Center Tissue and body fluid bank. All samples had been procured from consenting patients according to protocols approved by the Institutional Review Board and stored frozen. None of the samples had been thawed more than twice.

Triplicate serum samples ($n = 68$) from healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$), and HAM ($n_3 = 11$) patients were processed. A bioprocessor, which holds 12 chips in place, was used to process 96 samples at one time. Each chip contained one "QC spot" from normal pooled serum, which was applied to each chip along with the test samples in a random fashion. The QC spots served as quality control for assay and chip variability. The samples were blinded for the technicians who processed the samples. The reproducibility of the SELDI spectra, that is, mass and intensity from array to array on a single chip (intra-assay) and between chips (interassay), was determined with the pooled normal serum QC sample (Figure 1).

SELDI mass spectrometry

Serum samples were analyzed by SELDI mass spectrometry as described earlier [16]. The spectral data generated was used in this study for the development of the novel FDA.

Hierarchical clustering using functional data analysis

We propose to implement a hierarchical clustering algorithm for proteomics data using FDA, which consists of detecting hidden group structures within a functional dataset. We apply a new dissimilarity measure to the smoothed (transformed) proteomics functions $\hat{\mu}_i$. Then

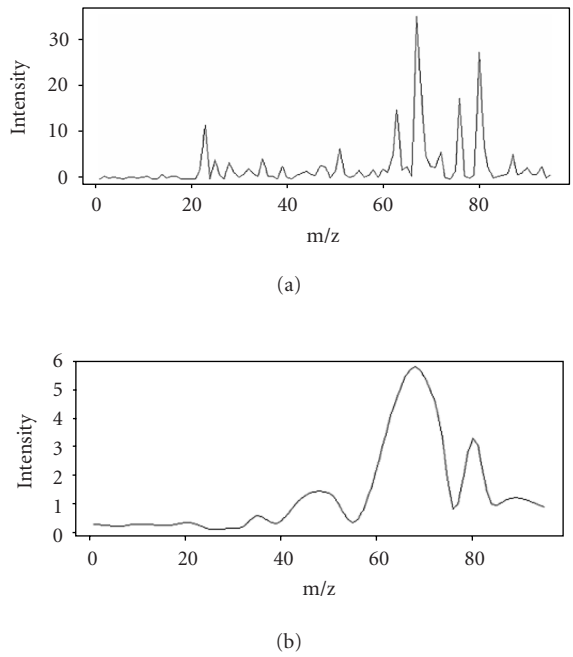


Figure 2. Original curve and a smoothed curve.

we develop a new metric that calculates the dissimilarity between different curves produced by protein expression. The development of metrics for curve and time-series models was first addressed by Piccolo [17] and Corduas [18]. Heckman and Zamar proposed a dissimilarity measure δ_{HZ} for clustering curves [19]. Their dissimilarity measure considers curve invariance under monotone transformations. Let $\Lambda_i = \{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\}$ be the collection of the estimated points where the curve $\mu_i(t)$ has a local maximum and let m_i be the number of maximals per observation or per sample $(i) \cdot \delta_{HZ}$ is defined as

$$\delta_{HZ}(i,l) = \frac{\sum_{j=1}^{m_i}\left(r\left(\lambda_j^{(i)}\right) - \overline{r(\lambda^{(i)})}\right)\left(r\left(\lambda_j^{(l)}\right) - \overline{r(\lambda^{(l)})}\right)}{\sum_{j=1}^{m_i}\left(r\left(\lambda_j^{(i)}\right) - \overline{r(\lambda^{(i)})}\right)^2 \sum_{j=1}^{m_l}\left(r(\lambda^{(l)}) - \overline{r(\lambda^{(l)})}\right)^2}, \tag{2}$$

where

$$r\left(\lambda_j^{(i)}\right) = k_j^{(i)} + \frac{u_j^{(i)}}{2}, \quad k_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} < \lambda_j^{(i)}\right\},$$
$$u_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} = \lambda_j^{(i)}\right\}, \qquad \overline{r(\lambda^{(i)})} = \frac{1}{m_i}\sum_{j=1}^{m_i} r\left(\lambda_j^{(i)}\right). \tag{3}$$

This measure is powerful for regression curves which are mainly monotone. On the other hand, Cerioli et al [20] propose a dissimilarity measure δ_C extending the one proposed by Ingrassia et al [21]. Cerioli's dissimilarity δ_C is defined by

$$d(i,l) = \sum_{j=1}^{m_i} \frac{|\lambda_j^{(i)} - \lambda_{*j}^{(l)}|}{m_i},$$
$$\lambda_{*j}^{(l)} = \left\{\lambda_{j'}^{(l)} : |\lambda_j^{(i)} - \lambda_{j'}^{(l)}| = \min,\ i = 1, \ldots, n\right\}, \tag{4}$$
$$\delta_C(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right).$$

Both dissimilarity measures show good performance for time-series data. Dissimilarity δ_C does not involve all the indices m_i of the smoothed curve. It also uses the shortest distance between curves by involving few data points obtained by FDA smoothing.

A flexible dissimilarity measure is the one that may combine the characteristic of both measures δ_{HZ} and δ_C. This means that a potential dissimilarity measure should use the collected estimated points of the original curve obtained from FDA so that no information is lost and should work on different type of smoothed curves without using the monotonicity restriction.

In this sense, we propose a functional-based dissimilarity δ_B measure which uses the rank of the curve proposed by Heckman and Zamar and generalizes Cerioli et al dissimilarity measure as follows:

$$d_{il} = \sum_{j=1}^{m_i} \frac{\left|r\left(\lambda_j^{(i)}\right) - r\left(\lambda_{*j}^{(l)}\right)\right|}{m_i},$$
$$r\left(\lambda_{*j}^{(l)}\right) = \frac{\sum_{h=1}^{m_l}\left|r\left(\lambda_j^{(i)}\right) - r\left(\lambda_{h'}^{(l)}\right)\right|}{m_l},$$
$$r\left(\lambda_j^{(i)}\right) = k_j^{(i)} + \frac{u_j^{(i)}}{2}, \quad k_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} < \lambda_j^{(i)}\right\}, \tag{5}$$
$$u_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} = \lambda_j^{(i)}\right\}, \qquad \overline{r(\lambda^{(i)})} = \frac{1}{m_i}\sum_{j=1}^{m_i} r\left(\lambda_j^{(i)}\right).$$

Obviously, $d_{ii} = 0$ and $d_{il} = 0$, if μ_i and μ_l have the same shape $(T_i = T_l)$. We can adjust the formula above to obtain a dissimilarity measure that satisfies symmetry, by taking δ_B as our proposed dissimilarity measure:

$$\delta_B(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right). \tag{6}$$

We used three powerful hierarchical methods to derive clusters or patterns using δ_B and we compare the performance of δ_B to δ_C and δ_{HZ}. The hierarchical algorithms we used are (1) *Pam* which partitions the data into different clusters "around their medoids," (2) *Clara* which works as in "Pam." Once the number of clusters is specified and representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid [22]. The sum of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the sum is minimal, is retained. Each sub-dataset is forced to contain the medoids obtained

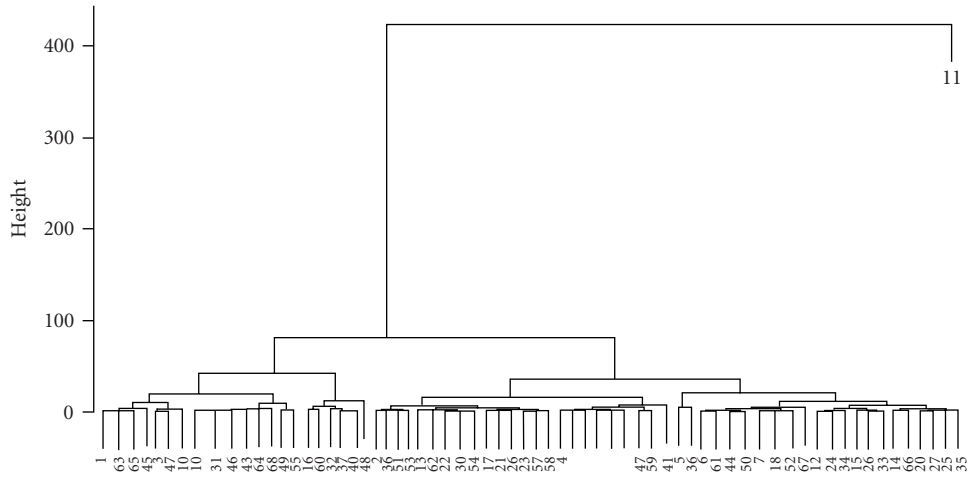


FIGURE 3. Clustering proteomics data with Diana.

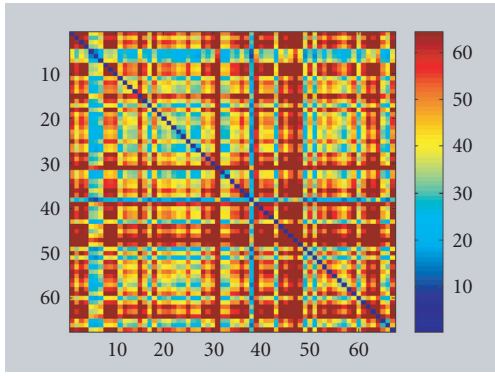


FIGURE 4. Pattern recognition using dissimilarity matrix δ_C.

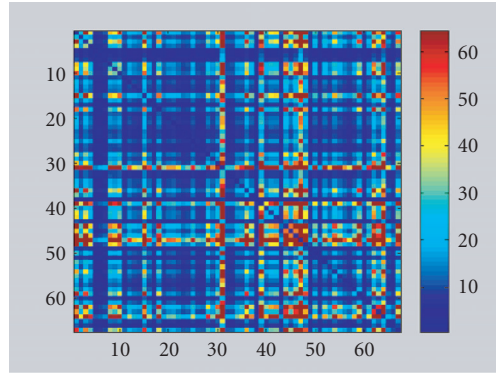


FIGURE 5. Pattern recognition using δ_{HZ}.

from the best sub-dataset until then. (3) *Diana* is probably unique in computing a divisive hierarchy, whereas most other software for hierarchical clustering is agglomerative. Moreover, Diana provides the divisive coefficient which measures the amount of clustering structure found. The Diana-algorithm constructs a hierarchy of clustering starting with one large cluster containing all n observations. Clusters are divided until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected [22].

RESULTS

Functional data transformation reduces the dimensionality of the spectra

The spectral data were collected from proteomics analysis of a total number of serum samples ($n = 68$) including healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$), and HAM ($n_3 = 11$) patients. The dataset is represented by an $n \times p$ matrix $\mathbf{X}$, where $p = 25,196$ is the number of variables (peaks) measured on each sample and $n = 68$ is the number of samples (patients). Any clustering algorithm on a datum ($68 \times 25,196$) will fail because of the singularity of the covariance matrix ($n < p$) and it will be difficult in manipulating matrices with 68 rows and 25,196 columns which has 1.7133×10^6 elements. This

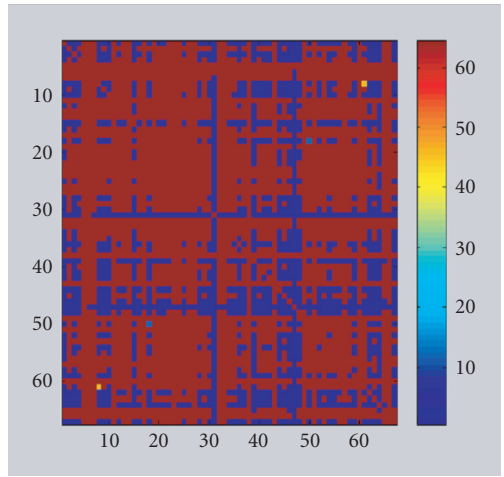


FIGURE 6. Pattern recognition using δ_B.

problem would not be raised for heuristic-based (ie, pairwise similarity-based) clustering algorithms.

To reduce the dimensionality of the spectral data, we applied FDA by fitting a P-spline curve $\widehat{\mu}_i(t)$ to each sample $\mathbf{y}_i$. P-splines satisfy a penalized residual sum of squares criterion, where the penalty involves a specified degree of derivation for $\mu_i(t)$. For example, cubic splines functions are P-splines of second order, penalizing the second derivative of $\mu_i(t)$. P-splines curves of order 3 penalize the third derivative of $\mu_i(t)$. P-splines curves of order 4 lead to an estimate of $\mu_i(t)$ with continuous first and second derivatives. We choose here to fit a P-spline curve of order 4 (Figure 2). The fitting step is performed by fixing the number of degrees of freedom that are implicit in the smoothing procedure [23].

The next step performed on the smoothed curves is to find the landmarks or indices T_i. We collected the first derivative of $\widehat{\mu}_i(t)$, say $\widehat{\mu}_i^{'}(t)$, using a smoothing P-spline function available in *R*. Those derivatives are crucial at determining the cut-off points or indices of $\mu_i(t)$. We performed this step by computing an approximate 95% pointwise confidence interval for the first derivative of $\mu_i(t)$ [24]. When the lower limit of this interval is positive, we have the confidence that $\mu_i(t)$ will be increasing. When the upper limit of this interval is negative, we have the confidence that $\mu_i(t)$ will be decreasing. Inside the interval, when the derivative changes from negative to positive, we have an optimal value which is a minimum. When the derivative changes from positive to negative, we have an optimal value which is a maximum. The maximum is set, for convenience, as the largest value of $\widehat{\mu}_i^{'}(t)$ in that interval. In this study, we restricted the choice of indices to maximal values. Let $\Lambda_i = \{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\}$ be the collection of the estimated points where the curve $\mu_i(t)$ has a local maximum and let m_i be the number of maximals per observation or per sample (i). Consequently, dissimilarity measure is calculated to derive the dissimilarity matrices of size $(n \times n)$ for all samples using the maximum values.

Clustering spectral data using functional data analysis

The application of functional data transformation led to the reduction of the dimensionality of the spectra to half. The size of mass indices become 12, 598. To cluster the reduced data, we calculated the three dissimilarity matrices M_{δ_C}, M_{δ_B}, and $M_{\delta_{HZ}}$. It appears that an unusual sample (patient 11) hides a possible pattern that we are trying to discover. Figure 3 shows a clustering dendrogram of the data using Diana approach. Pam and Clara gave the same results. This suggests that sample 11 would be important for further investigation.

When we removed observation 11, we detected a fewer fuzzy patterns with δ_C (Figure 4), δ_{HZ} (Figure 5), and δ_B (Figure 6). To be more specific, we investigated clusters proposed by δ_C and δ_{HZ}. A large number of clusters were proposed by both approaches (about 10 clusters). This strange result might be caused by the monotonicity assumption when using δ_{HZ} or the loss of information when using δ_C.

Table 1. Confusion matrix to show the performance of δ_B using Diana.

		Predicted			
	Classification	HAM	ATL	NOR	Total
Clinical	HAM	8	3	0	11
	ATL	5	14	1	20
	NOR	1	2	34	37
	Classification rate	0.73	0.70	0.92	0.84

Table 2. Confusion matrix to show the performance of δ_B using Clara.

		Predicted			
	Classification	HAM	ATL	NOR	Total
Clinical	HAM	10	1	0	11
	ATL	2	18	0	20
	NOR	1	1	35	37
	Classification rate	0.91	0.90	0.95	0.93

For δ_B, we provided the dendogram of the data using Diana approach (Figure 7). Three clusters were apparent. One well-separated cluster and two overlapped ones. For δ_{HZ} and δ_C, no structure was apparent which confirms the limitations of both dissimilarities as explained before.

To check the performance of our method, we calculated the confusion matrix between the predicted clusters and the clinical clusters using Diana (Table 1) and Clara (Table 2). We find that 3 patients out of 11 were misclassified for cluster 1 (HAM), 6 out of 20 were misclassified for cluster 2 (ATL), and 3 out of 37 were misclassified for cluster 3 (normal). Ham and ATL shared the majority of the misclassified observations which makes sense since both groups gather patients with a disease caused by the same retrospective virus. The error rate of misclassification for both clusters (HAM and ATL) is about 20%. For normal patient, the error rate of misclassification is about 8%. The total rate of misclassification is about 16%.

When we used Clara-based hierarchical cluster algorithm with δ_B, the classification result has dramatically been improved (Figure 8). The error rate of misclassification is reduced to 7%.The error rate of misclassification between HAM and ATL is about 9%, 5% of normal patients was misclassified. This result shows that a hierarchical δ_B dissimilarity algorithm based on minimizing the dissimilarity of observations to their closest medoid performs better than a divisive hierarchical clustering algorithm based on δ_B.

DISCUSSION

Cancer biomarkers can be used to screen asymptomatic individuals in the population, assist diagnosis in

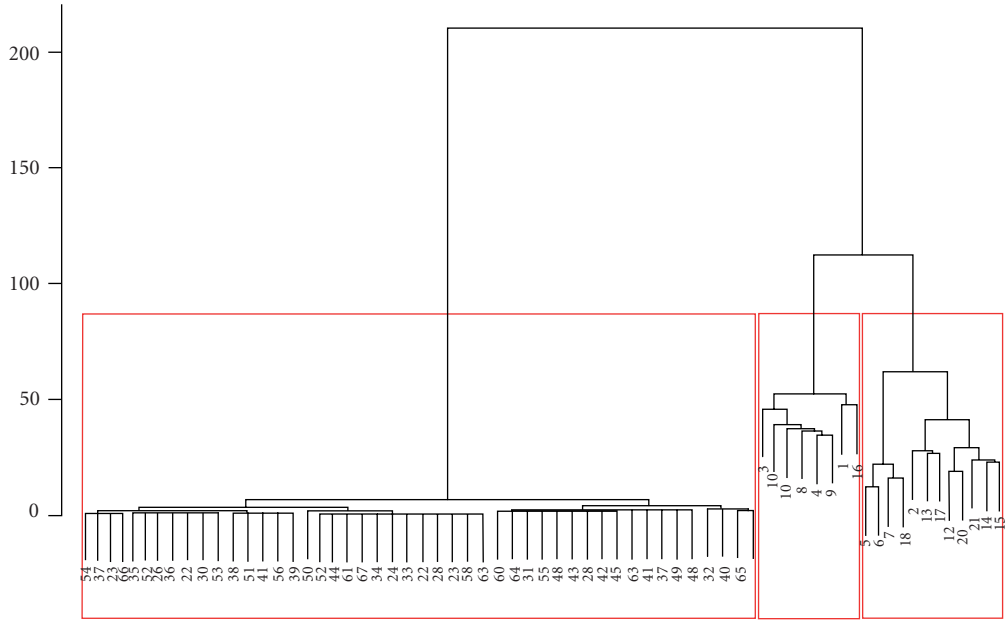


FIGURE 7. Dendogram of the δ_B dissimilarity approach with Diana.

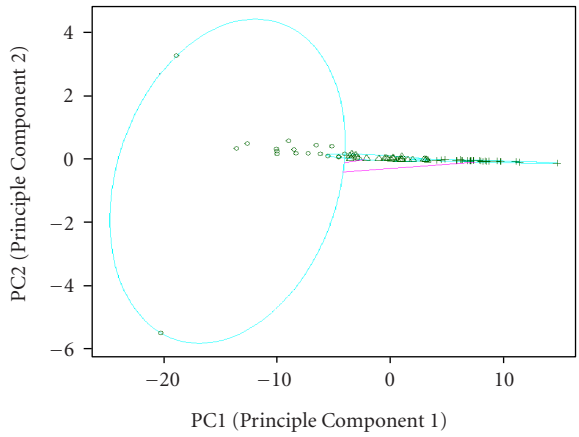


FIGURE 8. The δ_B dissimilarity approach with Clara.

suspected cases, predict prognosis and response to specific treatments, and monitor patients after primary therapy. The introduction of new technologies to the proteome analysis field, such as mass spectrometry, have sparked new interest in cancer biomarkers allowing for more effective diagnosis of cancer by using complex proteomic patterns or for better classification of cancers, based on molecular signatures, respectively. These technologies provide wealth of information and rapidly generate large quantities of data.

Processing the large amounts of data will lead to useful predictive mathematical descriptions of biological systems which will permit rapid identification of novel therapeutic targets and diseases biomarkers.

Clustering and analyzing proteomics data has been proven to be a challenging task.

Proteomics data are provided usually as curves or spectra with thousand of peaks. A clustering algorithm based on a matrix of n observations (n samples which is usually small) and p peaks (p variables which is usually a large number) will be unsuccessful. A matrix of size ($n \ll p$) will be singular and any method based on a matrix M ($n \times p$) will not be robust enough and will induce errors. A clustering algorithm based on a well-chosen dissimilarity matrix ($n \times n$) is more appropriate and more robust given the relatively moderate size of the matrix.

The use of a smoothing function for the spectra performs better for time series or for monotonic curves. We have previously successfully applied this smoothing function to large-scale proteomics data [25].

The application of Euclidean or Mahalanobis distances for instance may not perform well for this proteomics dataset, since those distances usually successfully applied to a typical data with specific expression, spherical or ellipsoidal (normally distributed data). A new dissimilarity measure has to involve other criteria such as the wealth of data points for each observation and the parallel nature expressed by the proteomics curve (or time series). On the other hand, a robust dissimilarity measure may perform badly on a curve with large data points or peaks.

Functional smoothing of proteomics expression profiles or spectra has proven to be very helpful. This has allowed us to minimize the number of peaks to retain only the ones that passed the performance of the FDA smoothing. In this study, after using FDA, we succeeded in retaining 50% of the smoothed peaks. The FDA with

the dissimilarity measure δ_B shows better performance by comparison to δ_C and δ_{HZ} known to perform well along with FDA on times-series data or on monotonic curves.

The two remaining difficulties that naturally arose are (1) to find meaningful peaks that can be used to provide better discrimination between the clusters, (2) to propose the optimal number of clusters instead of choosing them a priori. The model selection criteria might be useful to answer those questions. In fact, model selection scores use two components for selecting the number of variables and the number of clusters in a given density-based cluster analysis. The first term is the lack of fit generally proportional to the likelihood function. The second term is the penalty term (complexity term). For such proteomics dataset, we propose to use the sum of the negative δ_B dissimilarity measure between all the observations to their closest medoids as a lack of fit function. The penalty term might be simple to derive but biased using AIC and BIC, for example, or it can be more difficult to derive if one used a more robust method such as information complexity-based criteria.

ACKNOWLEDGMENT

This work was supported by the SRGP Award by the College of Business, University of Tennessee in Knoxville, by the Leukemia Lymphoma Society, and the National Institutes of Health.

REFERENCES

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198–207.

[2] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. 2004;5(9):699–711.

[3] Wright Jr GL. SELDI proteinchip MS: a platform for biomarker discovery and cancer diagnosis. *Expert Rev Mol Diagn*. 2002;2(6):549–563.

[4] Reddy G, Dalmasso EA. SELDI protein chip(R) array technology: protein-based predictive medicine and drug discovery applications. *J Biomed Biotechnol*. 2003;2003(4):237–241.

[5] Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev*. 2004;23(1):34–44.

[6] Espina V, Mehta AI, Winters ME, et al. Protein microarrays: molecular profiling technologies for clinical specimens. *Proteomics*. 2003;3(11):2091–2100.

[7] Zhang H, Yan W, Aebersold R. Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr Opin Chem Biol*. 2004;8(1):66–75.

[8] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*. 2003;21(1):697–700.

[9] Bensmail H, Haoudi A. Postgenomics: proteomics and bioinformatics in cancer research. *J Biomed Biotechnol*. 2003;2003(4):217–230.

[10] Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*. 2003;19(12):1484–1491.

[11] Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM. Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin Cancer Res*. 2004;10(3):981–987.

[12] Ramsay JO, Silverman BW. *Functional Data Analysis*. New York, NY: Springer; 1997.

[13] Ramsay JO, Silverman BW. *Applied Functional Data Analysis: Methods and Case Studies*. New York, NY: Springer; 2002.

[14] Haoudi A, Semmes OJ. The HTLV-1 tax oncoprotein attenuates DNA damage induced G1 arrest and enhances apoptosis in p53 null cells. *Virology*. 2003;305(2):229–239.

[15] Haoudi A, Daniels RC, Wong E, Kupfer G, Semmes OJ. Human T-cell leukemia virus-I tax oncoprotein functionally targets a subnuclear complex involved in cellular DNA damage-response. *J Biol Chem*. 2003;278(39):37736–37744.

[16] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*. 2002;62(13):3609–3614.

[17] Piccolo D. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*. 1990;11:153–164.

[18] Corduas M. La metrica autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS. *Quaderni di statistica*. 2000;2:1–37.

[19] Heckman N, Zamar R. Comparing the shapes of regression function. *Biometrika*. 2000;87(1):135–144.

[20] Cerioli A, Laurini F, Corbellini A. Functional cluster analysis of financial time series. In: *Proceedings of the Meeting of Classification and Data Analysis Group of the Italian Statistical Society (CLADAG 2003)*. Bologna, Italy: CLUEB; 2003:107–110.

[21] Ingrassia S, Cerioli A, Corbellini A. Some issues on clustering of functional data. In: Schader M, Gaul W, Vichi M, eds. *Between Data Science and Applied Data Analysis*. Berlin, Germany: Springer; 2003:49–56.

[22] Kaufman L, Rousseeuw PJ. *Finding Groups in Data. An Introduction to Cluster Analysis*. New York, NY: John Wiley & Sons; 1990.

[23] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. London UK: Chapman & Hall; 1990.

[24] Silverman BW. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J Roy Statist Soc B*.1985;47:1–52.

[25] Bensmail H, Semmens J, Haoudi A. Bayesian fast-Fourier transform based clustering method for proteomics data. *Journal of Bioinformatics*. In press.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 87–95 • DOI: 10.1155/JBB.2005.87

RESEARCH ARTICLE

Objective Clustering of Proteins Based on Subcellular Location Patterns

Xiang Chen[1,3] and Robert F. Murphy[1,2,3]

[1]*Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA*
[2]*Department of Biomedical Engineering, 2100 Doherty Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA*
[3]*Center for Automated Learning and Discovery, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891, USA*

Received 8 September 2004; accepted 4 November 2004

The goal of proteomics is the complete characterization of all proteins. Efforts to characterize subcellular location have been limited to assigning proteins to general categories of organelles. We have previously designed numerical features to describe location patterns in microscope images and developed automated classifiers that distinguish major subcellular patterns with high accuracy (including patterns not distinguishable by visual examination). The results suggest the feasibility of automatically determining which proteins share a single location pattern in a given cell type. We describe an automated method that selects the best feature set to describe images for a given collection of proteins and constructs an effective partitioning of the proteins by location. An example for a limited protein set is presented. As additional data become available, this approach can produce for the first time an objective systematics for protein location and provide an important starting point for discovering sequence motifs that determine localization.

INTRODUCTION

The biotechnology revolution, especially the development of high-throughput technologies, has led to a rapid explosion of biological raw data that could not been imagined a few decades ago. For the first time in history, biologists can perform metaanalysis on available experimental data (largely unorganized) in order to generate hypotheses for the mechanisms by which cells, tissues, and organisms carry out their specialized functions. Until recently, systematic efforts to describe protein location have been limited to the assignment by database curators of a relatively small set of terms to each protein. While the recent development of restricted vocabularies for this purpose (most prominently the Gene Ontology Consortium cellular component ontology) has been an important step, such vocabularies do not have the ability to uniquely identify the many (probably on the order of a hundred) distinct, complex subcellular patterns displayed by proteins. To complement these approaches, we have applied pattern recognition and machine learning methods to this general problem, and coined the term "location proteomics" to describe the branch of proteomics that systematically and objectively studies the location patterns of individual proteins and their relationships [1].

Cells vary greatly in their size, shape, intensity, position and orientationation in fluorescent images, and consequently raw pixel intensity values are not very useful in location pattern recognition in general. The core of our group's previous work has been the development of sets of numerical features (termed subcellular location features, or SLFs) to represent the patterns of proteins seen in fluorescence microscope images without being overly sensitive to changes in intensity, rotation, and position of a cell [2, 3]. These numerical descriptions of subcellular location have been validated by developing automated classifiers that can correctly assign previously unseen images to the major classes of subcellular structures or organelles [2, 3, 4, 5].

With the development of automated high-resolution microscopy technology [6, 7], the capability now exists for capturing high-resolution 3D fluorescence microscope images of protein subcellular distributions. Coupled with technologies that create cell lines expressing randomly tagged proteins [8, 9], it is possible to collect large numbers of images for diverse proteins in a given cell type

Correspondence and reprint requests to Robert F. Murphy, Department of Biomedical Engineering, 2100 Doherty Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA, Email: murphy@cmu.edu

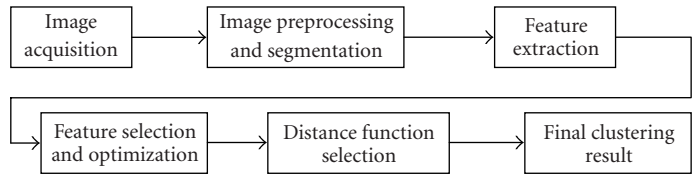


Figure 1. Flow chart for clustering protein subcellular location patterns.

within a reasonable time scale. The work described in this paper tries to approach the ultimate goal of determining which proteins imaged in such a project share the same location pattern. The problem could be stated alternatively as follows. *Given a set of proteins, each with multiple image representations, find a partitioning of the protein set such that images from members in the same partition show a single location pattern.*

From the computational view, the task of finding the optimal grouping of a set of proteins based on their subcellular location patterns can be described as finding the maximum number of partitions so that the SLF features of proteins within the same partition are statistically indistinguishable while the features for any two proteins from different partitions are distinguishable. This is the classic clustering problem. It is formally identical to those appearing in many other fields, such as identifying gene clusters from mRNA expression levels. In our case, however, image features are measured on widely varying scales with different units while mRNA expression levels are at least all expressed in the same units. This inhomogeneity of units complicates the process of feature selection and distance definition.

Building on our initial work demonstrating the use of the SLF to build subcellular location trees [1], we describe here clustering approaches for constructing objective partitionings of proteins by location. We started by making what we consider to be a reasonable assumption: that the majority of images for a specific protein in interphase cells should show the same location pattern. Under this assumption, we propose a method (shown as a flow chart in Figure 1) for automatically determining the number of partitions for the dataset and performing the partitioning accordingly.

METHODS

Image acquisition

A set of NIH 3T3 cells clones each expressing a different GFP-tagged protein were obtained by CD-tagging [10]. The acquisition of high resolution 3D images of these clones by spinning disk confocal microscopy has been described [1]. Briefly, the pixel spacing in both directions in the image plane was 0.11 μm and the vertical spacing between adjacent planes (slices) was 0.5 μm. The gray level of each pixel is between 0 and 4095 (12 bits per pixel). The resulting $1280 \times 1024 \times 31$ 3D images each contained from 1 to 3 cells. Ninety differently tagged protein clones (with 8 to 33 cells per clone) were included in the current study. Example images are shown in Figure 2.

Image preprocessing and segmentation

The procedures employed in image preprocessing have been described in detail previously [3, 11]. In brief, background in each image was removed and single-cell images were obtained through image segmentation (either automated or manual). Single-cell images were then thresholded using an automated method.

Feature extraction and optimization

The SLFs used in the current study can be divided into three categories:

(i) morphological features [5], based on finding the fluorescent objects in an image. A fluorescent object is a set of connected pixels with above threshold intensities,

(ii) edge features [1], which capture the amount of fluorescence distributed along edges,

(iii) haralick texture features [1, 12], based on the gray level co-occurrence matrix of an image, which capture the correlation between adjacent pixel intensities.

This combination of features (previously defined as 3D-SLF11) were extracted on the preprocessed single-cell images according to procedures described previously [1, 5, 12] with one modification. Previous studies suggested that pixel resolution and gray levels could potentially influence the discriminating power of the Haralick texture features [12, 13]. Therefore, the Haralick texture features were calculated at various degrees of downsampling (to 0.5, 1.0, 1.5, 2.0, and 2.5 μm pixel size) and various numbers of gray levels (16, 64, and 256). The optimal values for the current dataset were determined as 0.5 μm pixel size and 64 gray levels using the method described in [12].

Feature selection

For an arbitrary collection of proteins, it is not a trivial task to identify the optimal feature set to use for clustering them. One idea is to generate a number of different feature

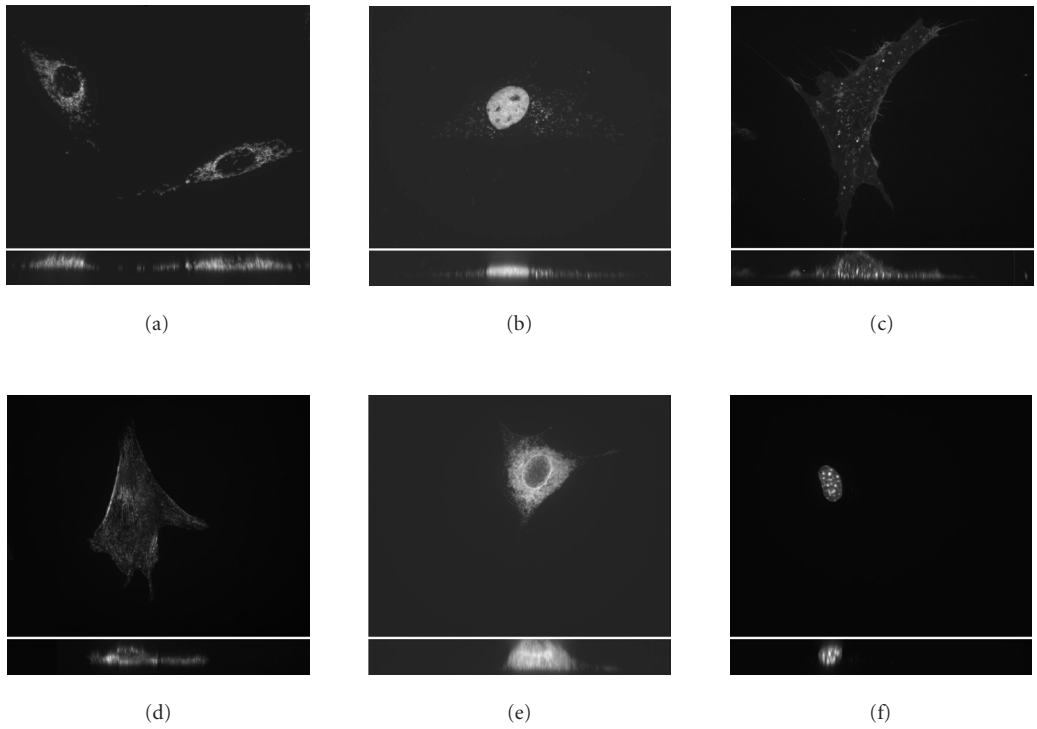


FIGURE 2. Selected images from the 3D 3T3 image dataset. Tagged protein names are shown with a hyphen followed by a clone number if the same protein was tagged in more than one clone in the dataset. Representative images are shown for (a) Atp5a1-1, (b) Ewsh, (c) Glut1, (d) Tubb2-1, (e) Canx, and (f) Hmga1-1. The top portion of each panel shows a projection on the x-y plane and the bottom shows a projection on the x-z plane.

sets and to use them to train a classifier that tries to distinguish every protein in the collection. We then consider the best feature set available to be the one with the highest overall classification accuracy. Of course, the overall accuracy is not an accurate estimate of the classifier's true discriminating power since some proteins in the collection may share a single location pattern. These proteins would be indistinguishable for a classifier and the classification result among these proteins could be largely random (lowering the overall accuracy). However, that accuracy is still a good metric for choosing a feature set since it will increase as informative features are added and decrease as they are removed.

Stepwise discriminant analysis (SDA) was used for selection of those "informative" features that support the discrimination between proteins with different patterns. The stepdisc function of SAS (SAS Institute, Cary, NC) was used with default parameter values (stepwise selection method, all variables included in the model calculation, start from no variable in the model, use 0.15 as the significance level for adding or retaining variables). The input to SDA was the full feature matrix for all cells for all clones and the output was a ranked list of features that were considered to contribute to distinguishing the clones.

To select the optimal feature subset for clustering, increasing numbers of the ranked features were used to train classifiers to try to distinguish each protein clone as described previously [12] with one exception: instead of the neural network classifier, we used a support vector machine (SVM) classifier with max-win strategy [14].

Distance function

As a starting point for this work, we used (1) a Euclidean distance function, which calculates the distance between each pair as the square root of the sum of squares of the feature differences over the whole feature set (each feature is normalized to zero mean and unit variance across the entire image collection) and (2) a Mahalanobis distance function, which further takes the correlations between features into account.

In the current 3D 3T3 dataset, most morphological, edge, and texture mean features were either in a single-mode, bell, shaped distribution (Figure 3a) or in a single-mode, exponential-shaped distribution (Figure 3b). The distributions for some texture range features were more complex with double modes (Figure 3c).

To avoid excessive weighting of features whose absolute values happen to be larger than the other features (compare Figures 3b and 3c) in Euclidean distance calculation, we first normalized all features to z-scores (subtracting the mean for each feature and dividing

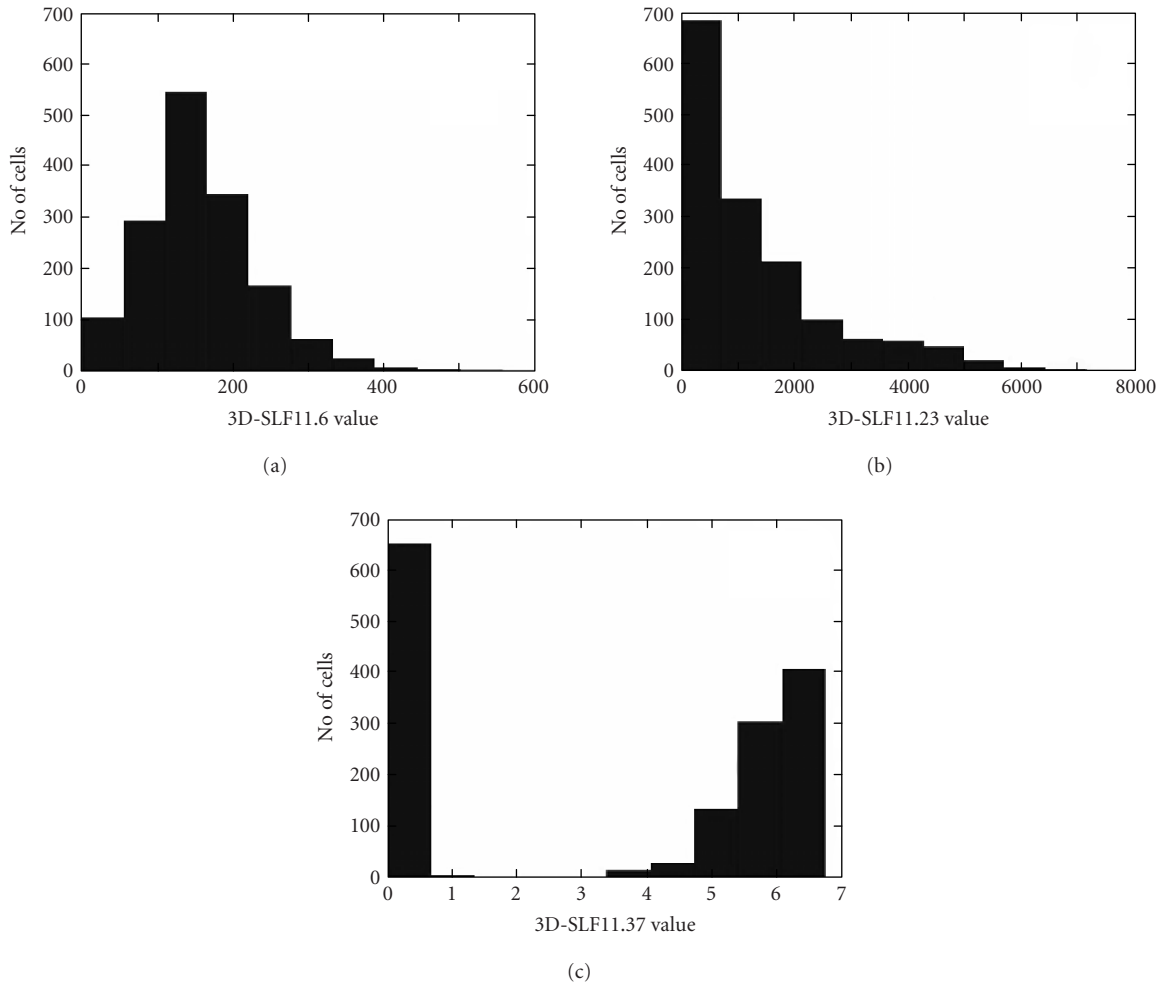


FIGURE 3. Histograms of selected features before z-score normalization. Examples of features with (a) a roughly Gaussian distribution (3D-SLF11.6, average object to center of fluorescence distance), (b) a roughly Poisson distribution (3D-SLF11.23, texture feature average of co-occurrence matrix sum variance), and (c) a biomodal distribution (3D-SLF11.37, texture feature range of co-occurrence matrix sum entropy).

by its standard deviation, both calculated across all clones).

Due to differences in the number of single-cell images for each clone, we randomly selected five images for each clone to construct a global covariance matrix. This process was repeated 100 times and the mean value was taken as the final covariance matrix used in the Mahalanobis distance calculation.

Clustering/partitioning algorithms

Each single-cell image from all clones was first converted to a feature vector and k-means clustering was performed on the entire image set using varying k (from 2 to the total number of clones). Akaike information content (AIC) was then used as a criterion to select the optimal value of k [15].

As a parallel approach, hierarchical clustering was performed on *mean* feature vectors for each clone. Since the image collection contains multiple images for each individual protein, we can construct many estimates of the mean feature vectors by randomly selecting half of the images for each protein. For each randomly chosen set, the simplest tree building algorithm, unweighted pair-group method with arithmetic mean (UPGMA) algorithm, was used to construct a distance tree (dendrogram). These were used to form a consensus tree [16] which contains those structures with general agreement in the set of trees for all random trials. Different partitionings of the protein set could be obtained by cutting the consensus tree at different heights (lower height yields more clusters). Optimal partitioning was selected using the AIC criteria.

A third approach (Algorithm 1) started with the confusion matrix created by the classifier described in the "feature selection" section. It is expected that some clones share a single location pattern. Consequently, images

from different clones with the same pattern will be expected to be assigned to one of those clones largely at random. To cluster these together, the confusion matrix was searched for off-diagonal elements that were above a threshold and the clones corresponding to these elements were merged. We select the threshold to yield a similar number of clusters to the optimal k obtained in the k-means/AIC algorithm.

The last approach we took was visual inspection where one or more descriptive term (e.g., uniform, cytoplasmic, nucleolar) was assigned for each clone after visually examining all images sequentially displayed on a monitor. Clones with the same combination of descriptive terms were grouped into the same cluster.

Evaluation of distance functions

The choice of distance function is critical to any clustering task. We do not intend to propose an optimal distance function here since it is possible that the best distance function should be determined individually for different datasets, either theoretically or experimentally. Instead we proposed to evaluate the effectiveness of different distance functions by measuring the agreement of the partitioning using the same distance function with different algorithms. The intuition behind this method is that a good distance function should be able to yield consistent partitioning of the dataset with different clustering algorithms.

In order to measure the degree of agreement among different clustering results, we used Cohen's κ statistic [17, 18] to compare two partitionings A and B:

$$\kappa(\mathrm{A},\mathrm{B}) = \frac{\text{Observed agreement} - \text{expected agreement}}{1 - \text{expected agreement}} \tag{1}$$

Observed agreement is defined as the portion of protein pairs where the two clustering results agree (the pair belongs either to a single cluster or to two distinct clusters for both results). Expected agreement can be defined as the agreement between two random partitionings with the same distribution frequencies as A and B, respectively. Calculating the expected agreement is difficult but it can be estimated by simulation. The κ statistic represents the portion of agreement in the two clustering results beyond chance, with a maximum value of 1 for perfect agreement. By running multiple simulations (randomly, independently assigning the set of clones to different partitions in A and B based on their marginal distribution probabilities and then calculating the observed κ statistic), it is also feasible to estimate the variance of the κ statistics under the null hypothesis that partitionings A and B are independently and randomly distributed.

RESULTS

The 3D 3T3 dataset, consisting of 90 randomly tagged protein clones, was obtained using CD-tagging techniques [8]. We first constructed the optimal feature subset to use in clustering these proteins by their location patterns, as described in the "Methods" section. Since morphological, edge, and texture features have all been shown to be useful for classifying both 2D [3] and 3D images [12], we began our search for discriminating features using a set of 42 features drawn from all three types. Using the method described before, a subset of 34 features (which we defined as 3D-SLF18, see Table 1) gave the best overall classification accuracy on a subset of 46 clones from the 3D 3T3 dataset (data not shown). This feature subset, consisting of 9 morphological features, 1 edge feature, and 24 texture features, was used for subsequent clustering procedures in this study.

We next consider approaches to clustering these proteins. As an initial approach, we propose clustering all individual images and determining an optimal number of clusters (the large number of individual images makes this estimate feasible). To do this, individual images were first converted to feature vectors and k-means clustering was then performed on the whole image set using various k values (from 2 to the total number of proteins included in the collection). Under the reasonable assumption that a majority of the cells in a clone share a single location pattern, the range of k should cover the optimal number of clusters/partitionings in the image set. Each value of k gave a specific clustering of the images with different cluster compactness (measured by the variances within the clusters). Akaike information content (AIC) was then used as a criterion to select the optimal value of k. AIC measures the fitness of the current model given the data, adjusted by the number of parameters included in the model (to avoid overfitting). This

```
Procedure clustering_on_confusionmatrix (Confusion-
Matrix, threshold)
Initialize cluster[i] = i for each i
While(max(off diagonal values in ConfusionMa-
trix) >= threshold) do
      normalize the ConfusionMatrix so that the
   sum of each row is 100
      select i < j such that ConfusionMatrix(i, j) is
    the largest above threshold off diagonal value of
    ConfusionMatrix
      set cluster[i] = cluster[i] ∪ cluster[j]
    clear cluster[j]
      set ConfusionMatrix[i,:] = ConfusionMa-
    trix[i,;] + ConfusionMatrix[j,:]
      set ConfusionMatrix[:,i] = ConfusionMa-
    trix[:,i] + ConfusionMatrix[:,j]
      clear ConfusionMatrix[j,:]
      clear ConfusionMatrix[:,j]
 End While
 return cluster
End Procedure
```

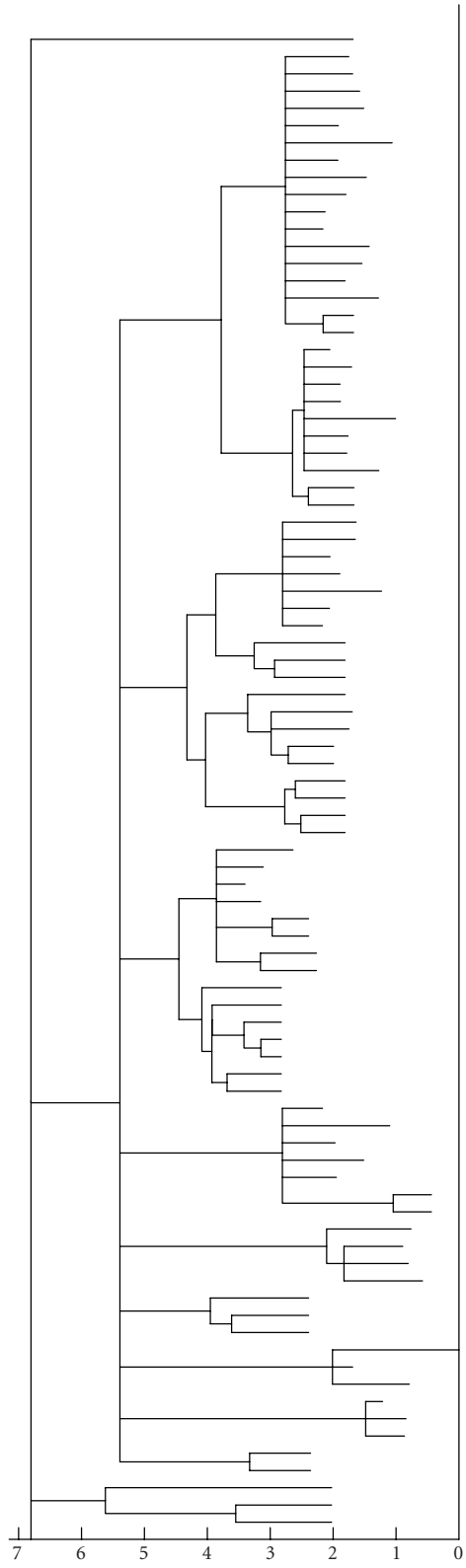
Algorithm 1. Procedure: clustering on confusionmatrix (ConfusionMatrix, threshold).

Table 1. Optimal feature set for distinguishing the 3D 3T3 images (3D-SLF18). The features are listed in decreasing order of discriminating power as evaluated by SDA.

Feature name	Feature description
3D-SLF11.16	The fraction of fluorescence in above threshold pixels that are along an edge
3D-SLF11.19	Average of correlation
3D-SLF11.23	Average of sum variance
3D-SLF11.31	Range of contrast
3D-SLF11.5	Ratio of maximum object volume to minimum object volume
3D-SLF11.28	Average of info measure of correlation 1
3D-SLF11.3	Average object volume (average number of above threshold pixels per object)
3D-SLF11.21	Average of inverse difference moment
3D-SLF11.24	Average of sum entropy
3D-SLF11.33	Range of sum of squares of variance
3D-SLF11.22	Average of sum average
3D-SLF11.29	Average of info measure of correlation 2
3D-SLF11.25	Average of entropy
3D-SLF11.34	Range of inverse difference moment
3D-SLF11.2	Euler number of the cell
3D-SLF11.41	Range of info measure of correlation 1
3D-SLF11.27	Average of difference entropy
3D-SLF11.26	Average of difference variance
3D-SLF11.37	Range of sum entropy
3D-SLF11.40	Range of difference entropy
3D-SLF11.35	Range of sum average
3D-SLF11.36	Range of sum variance
3D-SLF11.20	Average of sum of squares of variance
3D-SLF11.32	Range of correlation
3D-SLF11.4	Standard deviation (SD) of object volumes
3D-SLF11.38	Range of entropy
3D-SLF11.10	SD of absolute value of the horizontal component of object to protein center of fluorescence (COF) distances
3D-SLF11.9	Average absolute value of the horizontal component of object to COF distance
3D-SLF11.18	Average of contrast
3D-SLF11.13	SD of signed vertical component of object to protein center of fluorescence (COF) distances
3D-SLF11.6	Average object to COF distance
3D-SLF11.17	Average of angular second moment
3D-SLF11.42	Range of info measure of correlation 2
3D-SLF11.12	Average signed vertical component of object to protein center of fluorescence (COF) distances

gives a maximum likelihood estimate of the number of clusters given the data. Once the partitioning of the images was determined, all of the images belonging to the same protein were considered and the protein was allocated to the cluster that contained the maximum number of images from this protein as long as it accounted for at least 1/3 of the total images. Only those images belonging to this cluster were retained. When a given protein's images were found in several clusters so that none of the clusters had at least 1/3, that protein's location pattern was considered undetermined and it was dropped from further consideration. This reflects our initial assumption (or condition) that a protein has a unique pattern. The result of this stage is a clustering for only those protein images for which an assignment can be made with confidence.

As a parallel approach, we can perform hierarchical clustering on the average feature values for each protein (after eliminating the proteins considered too variable in the previous stage). We used the mean feature vector of each protein to construct a dendrogram. Since the image collection contains multiple images for each individual protein, we can construct many trees each of which is for a randomly selected half of the images for each protein. These are used to form a consensus tree [16], which contains the common structures with general agreement in



Unknown-4	Uniform	
Unknown-36	Cytoplasm + w nucleus	
Unknown-35	Cytoplasm + w nucleus	
Unknown-28	Cytoplasm	
Unknown-27	Cytoplasm	
Efl-gam	Cytoplasm	
Sep15	Cytoplasm	
Cnn2	Weak uniform	
Unknown-23	Cytoplasm	
Kars	Cytoplasm	Cytopl + unk
8430422M09Ri	Cytoplasm	
Rpl36	Cytoplasm	Rib + unk
Unknown-19	Weak uniform	
Unknown-18	Cytoplasm	
Txn1	Nuclmemb + w uniform	
Unknown-15	Cytoplasm	
Unknown-13	Cytoplasm	
Unknown-12	Cytoskeleton	
Unknown-34	Uniform	
Unknown-33	Uniform	
U17HG	Uniform	
Adfp-2	Uniform	
Unknown-24	Uniform	
Lgals1-2	Uniform	
Unknown-20	Uniform	
Unknown-14	Uniform	
Tpm4	Cytoplasm	
2610301D06Rik-2	Cytoplasm	Cytoplasm
Unknown-38	Cytoplasm	
Unknown-29	Small cytopl particle	
2700092o18Rik	Cytoplasm	
Unknown-17	Mito	
Unknown-16	Small cytopl part	
Rab21	Small cytopl part	Unk
Sdrp	Mito	
Atp5a1-2	Mito	Mito
Unknown-30	Small cytopl particle	
Adfp-1	Small cytopl particle	
**Ppar	Nucl + nuclmemb + w cyto	Nuc
Mrps18b	Mito	Mito + Rib + unk
Atp5a1-1	Mito	Mito
Timm23-5133400D	Small cytopl part	
Dia1	Small cytopl part	
NfiX-1	Nucleus + ER	Nuc + unk
Rtn3-1	ER	ER
**Rps6	Mito	Rib + cytopl + unk
Tubb2-1	Cytoskeleton	Cytoskeleton
Unknown-6	Cytoplasm	
Lasp1	Cytoplasm	
Rps11	Cytoplasm + nucleus	Rib + unk
2610301D06Rik-1	Cytoplasm	Cytoplasm
Tctex1	Cytoskeleton	Cytoskeleton
Sh3d3	Cytoskeleton	
Ewsh	Nucleus + w cytoplasm	Nucleus
**Hmgn2-2	Cytoplasm	Nucleus
Hmgn2-1	Nucleus	Nucleus
Canx	Cytopl + nuclmemb	ER
Unknown-1	Nucleus + w cytoplasm	
Unknown-11	Nucleus + w cytoplasm	
Similar to Siahbp	Nucleus + w cytoplasm	
Bat1a	Nucleus + w cytoplasm	Nucleus
Unknown-7	Cytoplasm + w nucleus	
Unknown-41	Cytoplasm	
Unknown-40	Cytoplasm	
Unknown-39	Nucl + cytopl + plasmemb	
Unknown-37	Cytoplasm	
Unknown-22	Cytoplasm	
Unknown-26	Sm cytopl part + w nucl	
Unknown-21	Cytoplasm + w nucleus	
Anxa2	Cytoplasm	
Unknown-3	Uniform	
Lgals1-1	Uniform	
Anxa5	Uniform	
Ltbp1-Pex12	Cytoplasm	
RP23-278K23	Cytoplasm	
**Prim2	Cytoplasm	Nucleus
Glut1	Cytoplasm + plasmemb	Unknown
Unknown-2	Cytoplasm	
Atox1	Mixture	
Unknown-9	Nucleus	
Hmga1-2	Nucleus	Nucleus
Hmga1-1	Nucleus	Nucleus
Unknown-5	Cytoplasm + w nucleus	
Ddx3	Cytoplasm + w nucleus	Unknown
**Rpl32	Nucleolar	Rib + unk
Unknown-32	Nucleolar	
Unknown-25	Nucleolar	

Figure 4. A consensus subcellular location tree generated on the 3D 3T3 image dataset using SDA-selected 3D-SLF11 features. The columns show the protein names (if known), human observations of subcellular location, and subcellular location inferred from gene ontology (GO) annotations. The sum of the lengths of horizontal edges connecting two proteins represents the distance between them in the feature space. Proteins for which the location described by human observation differs significantly from that inferred from GO annotations are marked (**).

Table 2. Comparison of clustering methods and distance functions. The agreement between the sets of clusters resulting from the four clustering methods described in the text was measured using the κ test. The standard deviations of the statistic under the null hypothesis were estimated to range between 0.014 and 0.023 from multiple simulations.

Clustering approaches compared	z-scored Euclidean distance	Mahalanobis distance
	κ	κ
k-means/AIC versus consensus	1	0.5397
k-means/AIC versus ConfMat	0.4171	0.3634
Consensus versus ConfMat	0.4171	0.1977
k-means/AIC versus visual	0.2055	0.1854
Consensus versus visual	0.2055	0.1156

the set of original trees. The clusters found in this tree can be compared to those obtained from clustering individual images.

We first compared the performance of the two different distance measures. It is reasonable to assume that a better distance function should produce greater agreement among clustering results using different algorithms. Since only the *k*-means/AIC and consensus hierarchical clustering algorithms utilized the distance function, we compared the agreement between the two clustering results using the κ statistic. In addition, we also compared these results against the results obtained using the other two algorithms (visual assignment and clustering using confusion matrix). Table 2 summarizes the results. Clearly the z-scored Euclidean distance function produced larger agreement than the Mahalanobis distance function. Therefore, we used the z-scored Euclidean distance function for the rest of the study. Another major point from Table 2 is that the agreements between visual clustering and the other approaches were clearly lower than the agreement between any pair of the machine clustering algorithms. The consistency seen among the automated methods confirms their value for generating location pattern annotations in proteomics projects.

When Euclidean distance was used as the distance function with the *k*-means/AIC algorithm for individual images, the optimal number of clusters found was 30. However, 13 of the 30 clusters contained only outliers from protein clones and therefore we obtained 17 clusters from this set of proteins. Out of all 90 clones, 3 were removed by the consistency requirement described above. The corresponding consensus tree obtained in parallel using average features is shown in Figure 4. The consensus tree was drawn in an additive style in which the sum of length of edges connecting pairs of proteins represents the distance between them.

Examination of the consensus tree (and the clusters obtained from *k*-means/AIC algorithms, not shown) reveals that proteins expected to have similar location patterns were mostly grouped properly. For example, the only three nucleolar proteins (Rpl32, Unknown-25, and Unknown-32) are grouped together. It should also be noted that there are two major nuclear protein clusters, one with Hmga-1, Hmga-2, and Unknown-9 and the other with Unknown-1, Unknown-11, similarly to Siahbp1 and Bat1a. The first cluster contained proteins with an exclusively nuclear distribution while the second cluster contained nuclear proteins with minor cytoplasmic distributions. The separation of these proteins into two clusters indicates that they are statistically distinguishable, in agreement with our previous results [1].

The consensus tree in Figure 4 has been incorporated into a web interface (http://murphylab.web.cmu.edu/services/PSLID/) that allows the underlying images for any branch to be displayed interactively.

CONCLUSIONS AND DISCUSSION

We have previously shown that the major protein subcellular location patterns can be described numerically by SLFs. Automated classifiers trained on these features can determine protein location patterns from previously unseen fluorescence images.

The observation that the SLFs used for this automated classification were clearly effective in distinguishing subcellular patterns suggested that a properly chosen partitioning of proteins using SLFs would group proteins based on their location patterns. We describe automated methods to create such a partitioning objectively. Our initial trial on a modest set of randomly tagged proteins using a set of morphological, edge, and texture features largely validates this method.

It should be pointed out that by increasing the dimensionality of protein images (e.g., by adding time as a fourth dimension and the presence of various drugs as a fifth dimension), proteins currently in the same cluster would be potentially distinguishable. This will of course require development of new features that reflect the characteristics of the higher dimensions.

In closing, we suggest that the development of an automated, systematic, and objective clustering approach for protein location patterns is critical to finding potential targeting motifs in protein sequences, just as automated clustering of gene expression data has been a prerequisite for automated detection of regulatory elements [19, 20, 21].

ACKNOWLEDGMENTS

We thank Jonathan Jarvik, Peter Berget, and all of our colleagues in the CD-tagging project for helpful discussions and providing images of tagged cell lines. This work was supported in part by NIH grant R01 GM068845, NSF grant EF-0331657, and a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund. X. Chen was supported by a graduate fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation.

REFERENCES

[1] Chen X, Velliste M, Weinstein S, Jarvik JW, Murphy RF. Location proteomics - building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc SPIE.* 2003;4962:298–306.

[2] Murphy RF, Boland MV, Velliste M. Towards a systematics for protein subcelluar location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:251–259.

[3] Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics.* 2001;17(12):1213–1223.

[4] Huang K, Murphy RF. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics.* 2004;5(1):78.

[5] Velliste M, Murphy RF. Automated determination of protein subcellular locations from 3d fluorescence microscope images. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*. New York, NY: IEEE; 2002:867–870.

[6] Nakano A. Spinning-disk confocal microscopy–a cutting-edge tool for imaging of membrane traffic. *Cell Struct Funct.* 2002;27(5):349–355.

[7] Price JH, Goodacre A, Hahn K, et al. Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J Cell Biochem Suppl.* 2002;39:194–210.

[8] Jarvik JW, Adler SA, Telmer CA, Subramaniam V, Lopez AJ. CD-tagging: a new approach to gene and protein discovery and analysis. *Biotechniques.* 1996;20(5):896–904.

[9] Rolls MM, Stein PA, Taylor SS, Ha E, McKeon F, Rapoport TA. A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J Cell Biol.* 1999;146(1):29–44.

[10] Jarvik JW, Fisher GW, Shi C, et al. In vivo functional proteomics: mammalian genome annotation using CD-tagging. *Biotechniques.* 2002;33(4):852–854, 856, 858–860 passim.

[11] Hu Y, Murphy RF. Automated interpretation of subcellular patterns from immunofluorescence microscopy. *J Immunol Methods.* 2004;290(1-2):93–105.

[12] Chen X, Murphy RF. Robust classification of subcellular location patterns in high resolution 3d fluorescence microscope images. *Proceeedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New York, NY: IEEE; 2004:1632–1635.

[13] Murphy RF, Velliste M, Porreca G. Robust classification of subcellular location patterns in fluorescence microscope images. *Proceeedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP 2002)*. New York, NY: IEEE;2002:67–76.

[14] Huang K, Murphy RF. Automated classification of subcellular patterns in multicell images without segmentation into single cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging (ISBI 2004)*. New York, NY: IEEE; 2004:1139–1142.

[15] Ichimura N. Robust clustering based on a maximum-likelihood method for estimating a suitable number of clusters. *Syst Comp Jpn.* 1997;28(1):10–23.

[16] Thorley JL, Page RM. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics.* 2000;16(5):486–487.

[17] Cook R. Kappa. In: P. Armitage and T. Colton, eds. *The Encyclopedia of Biostatistics.* New York, NY: Wiley; 1998:2160–2166.

[18] Cook R. Kappa and its dependence on marginal rates. In: P. Armitage and T. Colton, eds. *The Encyclopedia of Biostatistics*. New York, NY: Wiley; 1998:2166–2168.

[19] Jelinsky SA, Estep P, Church GM, Samson LD. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol.* 2000;20(21):8157–8167.

[20] Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx. *Curr Biol.* 2000;10(6):301–310.

[21] Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.* 2003;34(2):166–176.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 96–103 • DOI: 10.1155/JBB.2005.96

RESEARCH ARTICLE

High-Betweenness Proteins in the Yeast Protein Interaction Network

Maliackal Poulo Joy, Amy Brock, Donald E. Ingber, and Sui Huang

Vascular Biology Program, Departments of Surgery and Pathology, Children's Hospital, Harvard Medical School, Boston, MA 02115, USA

Received 1 June 2004; accepted 13 August 2004

Structural features found in biomolecular networks that are absent in random networks produced by simple algorithms can provide insight into the function and evolution of cell regulatory networks. Here we analyze "betweenness" of network nodes, a graph theoretical centrality measure, in the yeast protein interaction network. Proteins that have high betweenness, but low connectivity (degree), were found to be abundant in the yeast proteome. This finding is not explained by algorithms proposed to explain the scale-free property of protein interaction networks, where low-connectivity proteins also have low betweenness. These data suggest the existence of some modular organization of the network, and that the high-betweenness, low-connectivity proteins may act as important links between these modules. We found that proteins with high betweenness are more likely to be essential and that evolutionary age of proteins is positively correlated with betweenness. By comparing different models of genome evolution that generate scale-free networks, we show that rewiring of interactions via mutation is an important factor in the production of such proteins. The evolutionary and functional significance of these observations are discussed.

INTRODUCTION

The availability of genome-scale databases of pairwise protein interactions data in yeast [1] has made it possible to analyze the structure of the entire protein interaction network (PIN) in light of concepts from graph theory and the study of complex networks [2]. In these models of cell regulatory networks, proteins are represented by the nodes and the interactions between these components by the edges of the graph. Such genome-scale analysis of the PIN revealed that these molecular components form a "genome-wide" network, that is, the largest connected network component ("giant component") encompasses a dominant portion of the proteome. The large-scale topology (architecture) of this genome-wide PIN exhibits several interesting features that distinguish it from an Erdos-Renyi (ER) random graph [3]. For instance, the distribution of the connectivity (or degree, as used in graph theory) k which refers to the number of first neighbors of a given node approximates a power law, or, in other words, the PIN may be a scale-free network. PIN contains a larger number of highly connected proteins (hubs) than one would expect to find in an ER random network [4]. The connectivity of a protein appears to be positively correlated with its essentiality [4] in that highly connected proteins tend to be more essential for the viability of the organism.

Barabasi and Albert [5] proposed a simple algorithm for network growth (BA model) in which incoming nodes (newly evolved proteins) attach preferentially to existing nodes with higher degree. However, the yeast PIN exhibits additional structural details not observed in these randomly generated, scale-free networks. For instance, there are correlations between the connectivities of directly interacting proteins, in that connections between hubs are almost entirely absent. This feature has been postulated to be partly responsible for the robustness of biological networks [6]. Some specific local network structures, so-called network motifs, have also been shown to occur more frequently in molecular networks than in random networks [7]. Another structural feature of biological systems is their modularity, for example, the metabolic network exhibits a hierarchical modular structure [8].

In contrast to the genome-scale perspective, characterization of the biological functions of proteins has traditionally assumed the existence of distinct signaling modules that can be associated with particular cellular functions [9]. Hence, much effort has been spent

Correspondence and reprint requests to Sui Huang, Vascular Biology Program, Departments of Surgery and Pathology, Children's Hospital, Harvard Medical School, Boston, MA 02115, USA, E-mail: sui.huang@childrens.harvard.edu

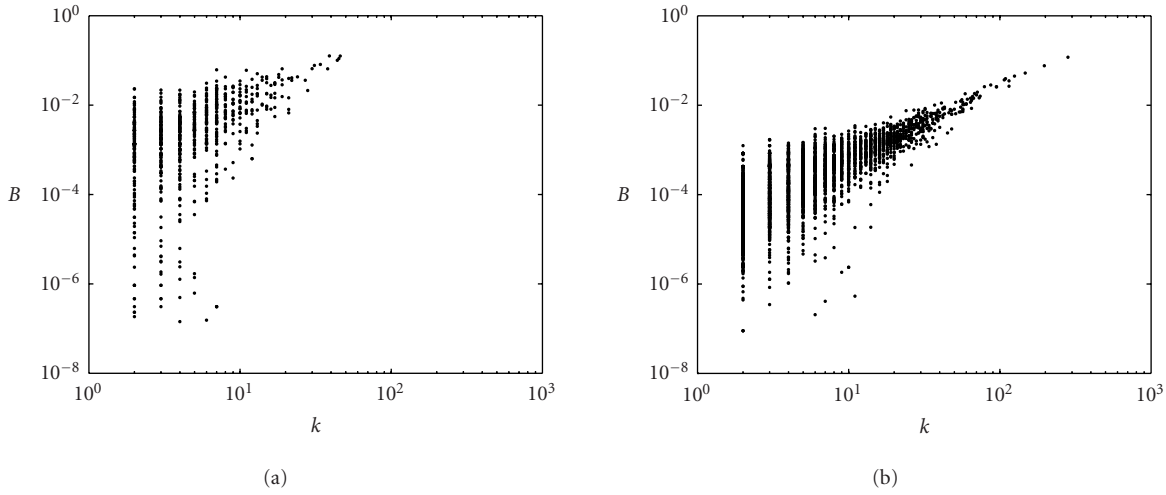


FIGURE 1. Degree (k) versus betweenness (B) plotted in logarithmic scale for the measured yeast interaction network based on DIP data [15, 16]. (a) Core data. (b) Full DIP data.

in defining and identifying discrete, functional network modules within the PIN. However, the ad hoc structural criteria used to define a module in physical networks remain somewhat arbitrary. Here we set out to examine a feature of complex networks that unites local and global topological properties of a node: the betweenness centrality. Measures, such as the connectivity of a node, k, and the clustering coefficients of networks, C [10], used previously to describe global architectural features capture only the local neighborhood of network nodes (nearest neighbors). In contrast, betweenness B_i of a given node i in a network is related to the number of times that node is a member of the set of shortest paths that connect all the pairs of nodes in the network (see "data and methods" for details). Hence, betweenness accounts for direct and indirect influences of proteins at distant network sites and hence it allows one to relate local network structure to global network topology [11]. Betweenness has also been used to characterize the "modularity" (eg, community structure) of various natural and man-made networks (see [12, 13]).

The functional relevance of the betweenness centrality B_i of a node is based on the observation that a node which is located on the shortest path between two other nodes has most influence over the "information transfer" between them. The betweenness distribution $P(B)$ of the nodes in a scale-free network also follows a power law or has a scale-free distribution, $P(B) \sim B^{-\rho}$ [14]. Although the distribution of the connectivity k across the nodes of the network has been used as a measure to characterize natural networks and the value of k has been suggested to correlate with the importance of the protein, this is truly valid only if the immediate neighbors are the only ones determining the properties of a protein in the network. In contrast, betweenness indicates how important the node is within the wider context of the entire network.

Based on analysis of the betweenness measure, we report here a new topological feature in the yeast PIN that is not found in randomly generated scale-free networks: the abundance of proteins characterized by high betweenness, yet low connectivity. The existence of such proteins points to the presence of modularity in the network, and suggests that these proteins may represent important connectors that link these putative modules. We describe here an extended network-generating algorithm that produces networks containing high betweenness nodes with low connectivity. We then discuss the evolutionary and functional significance of these findings.

RESULTS

We studied yeast protein interaction data obtained from different databases [1], including the Database of Interacting Proteins (DIP), and the Munich Information Center for Protein Sequences (MIPS) [15, 16, 17]. Although these networks differ at the level of individual protein-protein interactions, they exhibited the same global statistical properties. Here we present results for the most recent "full" [15] and "core" DIP data. In the core data only confirmed interactions were included [16]. The data set used contains 15 210 interactions between 4721 proteins for the "full" data set, and 6438 interactions among 2605 proteins for the "core" data set.

High-betweenness, low-connectivity proteins

Unlike the connectivity k which ranged from 1 to 282 in the PIN, values for betweenness B ranged over several orders of magnitude. The few highly connected nodes (hubs) in the PIN must have high-betweenness values because there are many nodes directly and exclusively connected to these hubs and the shortest path between these nodes goes through these hubs. However,

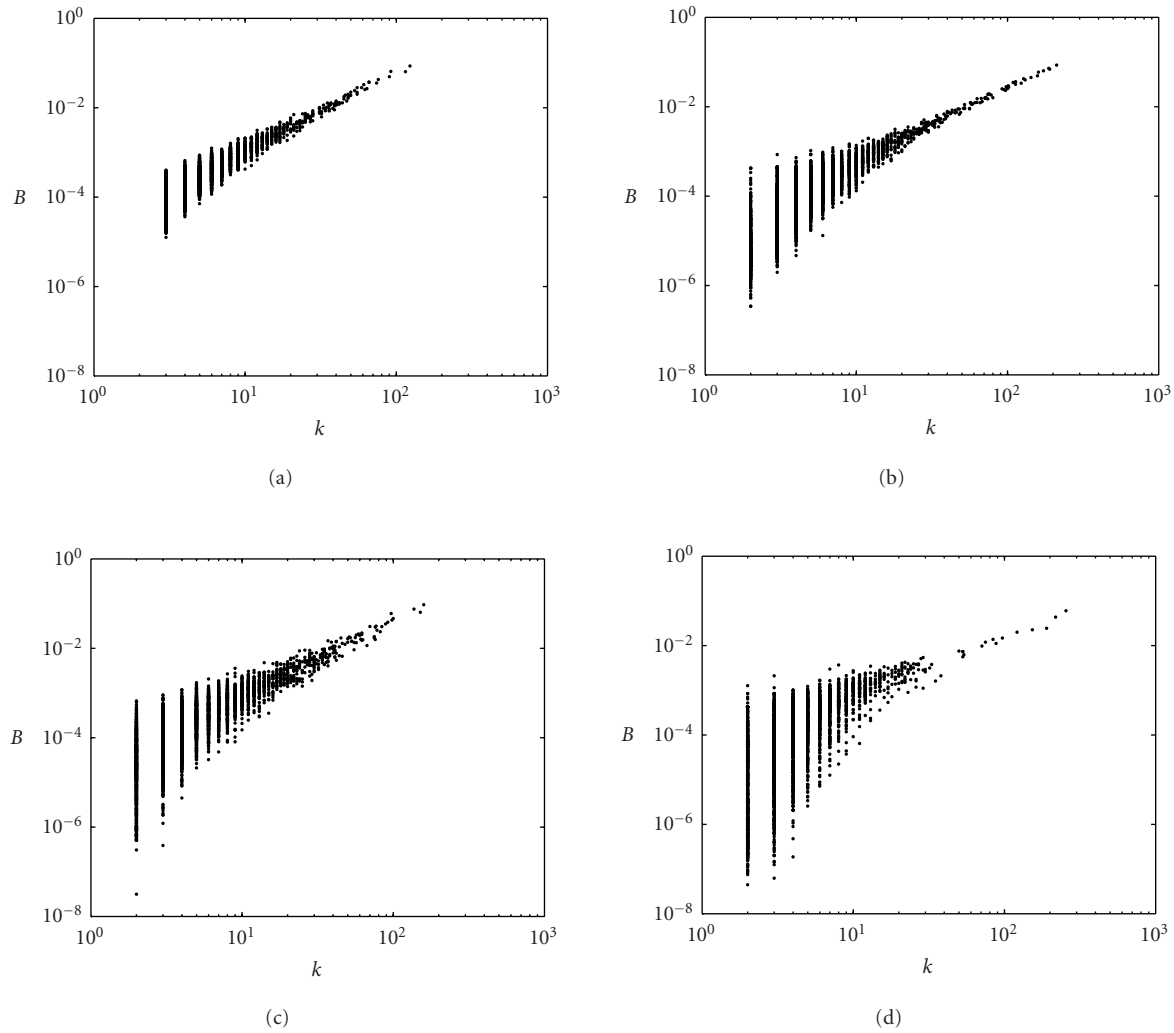


FIGURE 2. The $k - B$ plot for various model generative algorithms. (a) Barabasi-Albert (BA); (b) extended BA (EBA); (c) Sole-Vazquez (SV); (d) duplication mutation (DM).

the low-connectivity nodes also exhibited a wide range of betweenness values in the yeast PIN, as shown in Figure 1a (core data) and in Figure 1b (full data), where betweenness (B) is plotted as a function of connectivity (k). This indicates the existence of a large number of nodes with high betweenness but low connectivity (HBLC nodes). Importantly, such nodes are absent in computer-generated, random scale-free networks [5]. Although the low connectivity of these HBLC proteins would imply that they are unimportant, their high betweenness suggests that these proteins may have a global impact. From a topological point of view, HBLC proteins are positioned to connect regions of high clustering (containing hubs), even though they have low local connectivity.

Models

Can models for network evolution reproduce HBLC behavior? To address this question, we analyzed different computational models of biological network evolution that generate scale-free networks. The simplest generative algorithm, first proposed by Barabasi and Albert [5] (BA model) to explain the power-law distribution of connectivity, does not predict the existence of HBLC nodes: betweenness and connectivity were almost linearly correlated (Figure 2a). The extended Barabasi-Albert (EBA) model [18], where link addition and rewiring occur along with node addition with preferential attachment, also did not produce networks with HBLC nodes similar to that found in our analysis of the PIN, although low k nodes showed some spread of betweenness (Figure 2b). Moreover, this algorithm has no biological basis. A biologically motivated model put forward by Sole et al [19] and Vazquez et al [20] incorporated "gene duplication" as the driving mechanism for genome growth. In this model, the existing nodes (proteins) are copied with all their existing links, followed by divergence of the duplicated nodes introduced by rewiring and/or addition of connections, imitating mutations of duplicated genes. For the model

parameter range that produces power-law networks, the Sole-Vazquez (SV) model also failed to produce the same bias towards HBLC exhibited by the PIN (Figure 2c).

Berg et al (see [21]) have proposed a model that attempts to capture the actual molecular mechanism of genome growth based on evolutionary data. We asked whether that model can produce HBLC-node-containing networks. For our simulation of network growth, we used a modified version of the Berg model [21] which considered gene duplications and point mutations. "Duplications" relate to the process by which a gene is duplicated with all of its connections and which accounts for the increase in genome size, and hence network growth. "Point mutations" affect the structure of a protein such that it changes its interacting partners and hence connections within the network. The time scales involved in these two processes are different. Gene duplication is very slow compared to point mutation. The observed rate of gene duplication is less than 10^{-2} per million years per gene in *Saccharomyces cerevisiae*, while the point mutation rate is at least one order of magnitude higher [21]. Point mutations which affect a protein's ability to engage in molecular interactions are modeled as attachment or detachment of links, while the number of nodes is fixed ("link dynamics"). Since node duplication in evolutionary time scales is slow, compared to the time scale of link dynamics, gene duplication is modeled as addition of nodes without any links, while link dynamics occurs at each time step. This has been justified by the observation that in duplicated genes complete diversification occurs almost immediately after duplication. Usually, this divergence is biased, in that one of the proteins retains most of the interactions while the other retains a few or none [22]. Thus, for link dynamics in our simulation, a new attachment is established as follows: a random node is selected and attached to another node with preferential attachment, that is, with a rate proportional to its connectivity k as in the BA model. In contrast, for detachment, a link between two nodes is selected with a detachment rate proportional to the sum of inverses of their connectivities. This is motivated by the observation of higher mutation rates for less connected proteins [22, 23]. Importantly, simulation of network growth based on this duplication-mutation (DM) model led to the evolution of a network that exhibited power-law behavior with HBLC nodes (Figure 2d) similar to that exhibited by the yeast PIN. (See "data and methods" for details of model implementation.)

To compare the extent to which the various models produced HBLC nodes consistent with our experimental PIN data, we quantified the variation of betweenness values for a particular connectivity and its change with the value of the connectivity. In the basic BA network, betweenness and connectivity were almost linearly correlated in a logarithmic plot (Figure 2a). Thus, an increase in the standard deviation of betweenness values $D_B(k)$ among the nodes of a particular connectivity k, with decreasing k, reflects the presence of HBLC nodes. The plot of D_B versus the logarithm of k falls on a straight line, which will be flat if HBLC nodes are absent. The slope S of the best-fit straight line can thus be used as a measure for the presence of HBLC nodes (Figure 3). Our DM model had a slope very close to that of the PIN data while other models had significantly lower values of S.

Taken together, these results show that existing growth algorithms that produce scale-free networks do not predict the existence of HBLC nodes found within the yeast PIN. In contrast, a new model that is biologically more realistic, and considers mutations (random rewiring) in addition to duplication (node and link addition), produces a global network architecture with HBLC nodes that is consistent with the PIN of living cells. This finding supports the general idea that a trait, in this case, a network topology feature, may arise during evolution because of its inherent robustness due to mechanistic and historical constraints [24, 25]. However, it does not exclude contributions due to functional adaptation driven by natural selection, since the two mechanisms of genesis are not mutually exclusive.

Essentiality

Therefore, to address a possible role of selective pressure in the bias in betweenness in the PIN, we examined the relationship between a protein's essentiality and its betweenness value. Overall, we found that essential proteins of the yeast PIN had a higher mean betweenness and the frequency of high-betweenness nodes is greater for essential proteins. Mean betweenness for all proteins was 6.6×10^{-4} but for the essential proteins it was 1.2×10^{-3}; this represents an increase of 82%. In the case of connectivity, the increase of the connectivity value of essential proteins relative to all proteins was 77%. Thus, the betweenness of a protein reflects its essentiality to at least the same degree as its connectivity [4]. In Figure 4, the percentage of essential proteins among proteins within a particular range of betweenness values is displayed as a function of betweenness. The increase in the variance of betweenness values for low-connectivity proteins disrupts this correlation for low-connectivity values, whereas it does not disrupt the correlation between betweenness and essentiality. This is interesting, because HBLC proteins are not "protected" from mutation by the constraint imposed by a high number of interaction partners as in the case of high-connectivity nodes [23] and thus they could easily lose their betweenness property.

Evolutionary age

The association of essentiality with low connectivity embodied by the HBLC proteins raises the question about the relationship between betweenness and the evolutionary age of a protein. The BA model of preferential attachment would suggest that high-connectivity proteins, which are typically essential, evolved earlier, while low-connectivity proteins are more likely to be recent additions to the network [26]. To estimate the evolutionary age of proteins, we used the list of isotemporal categories

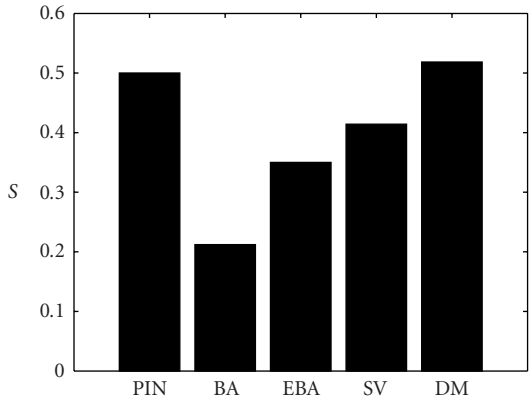


FIGURE 3. Magnitude of slope *S*, for the PIN data and different models (measured yeast protein interaction network (PIN) (data as in Figure 1b); the models are Barabasi-Albert (BA); extended BA (EBA); Sole-Vazquez (SV); duplication mutation (DM)). *S* measures the decrease of variance of the betweenness values of proteins with increasing degree, and hence indicates the relative prevalence of HBLC proteins.

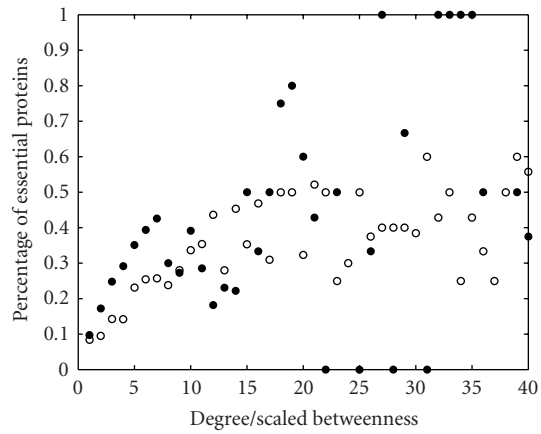


FIGURE 4. Percentage of essential genes with a particular degree (open circle) or betweenness (filled circle). Betweenness is scaled in such a way that the maximum value of betweenness is equal to the maximum degree. The plot was truncated at $k/B = 40$, since the number of essential genes beyond that is too small to have statistical significance.

of yeast protein orthologs provided by Qin et al [27], and classified them into four different age groups based on the phylogenetic tree, as in [26]. The core data set with confirmed interactions [16] showed a linear dependence of age and connectivity, while the dependence was not linear for the full data set [15], although there was a positive correlation (see Figure 5). The latter finding is consistent with the notion that some of the connections listed in the full data set are false positives [16].

Since betweenness correlates with essentiality and evolutionary age, it would be of particular interest to determine if the group of HBLC proteins has a different age or essentiality than the non-HBLC proteins of the same connectivity degree. Unfortunately, the number of proteins that falls into this class is too small to make statistically robust conclusions. This is because essentiality expressed as a continuous quantity as is done here and elsewhere [4, 26] is actually a group property (percentage of indispensable proteins in a given group) and not an attribute of individual proteins. Age is also a crude measure in that only four age groups can be defined; thus both measures require large numbers of proteins. With these caveats, our analyses found no statistically significant difference in evolutionary age or in essentiality between the HBLC proteins and their low-betweenness counterparts of the same connectivity.

DISCUSSION

Here, we report a new topology feature in the PIN not found in random networks: the prevalence of low-connectivity-degree nodes with high-betweenness values. It is also not predicted by the elementary growth model that explains the scale-free property of the PIN [5]. The existence of architectural features that deviate from that of a random graph immediately raises the fundamental question of how such a nonrandom network structure first originated. In general, one can distinguish two main mechanisms of genesis that can contribute to a particular biological, nonrandom feature: (i) adaptive evolution toward optimization of a function and (ii) inherent robustness due to constraints imposed by the particular history and mechanism of its formation [24, 25]. The former explanation, which represents Darwinian selection of the fittest, is equivalent to the engineer's notion of functional optimization. Its validation typically rests on the demonstration of convergent evolution and of a functional advantage. Thus, it requires analysis of the specific identity of the nominal proteins, their evolutionary (historical) relationships, as well as the phenotypic consequences of that network structure [28, 29, 30]. In contrast, inherent robustness due to network constraints is more fundamental and implies that a nonrandom feature is the unavoidable consequence of some elementary physical, mechanistic, or other less obvious, self-organizing principles [25, 31]. As for networks, this second mechanism can be reduced to a simple, generic, generative algorithm that may represent a plausible mechanism for the genesis of a given system, as has been studied by researchers in the field of complexity [31, 32, 33]. Hence, network structures are particularly well suited for addressing the relative contribution of either mechanism responsible for formation of a nonrandom trait [25].

By comparing network growth models, we found that mutation (changes in network links due to addition and deletion) is central to the mechanism of network genesis that produces HBLC nodes. Thus, our simple algorithm explains this network topology feature without invoking functional adaptation. In this study on the generic architecture of the PIN, we do not discuss the molecular

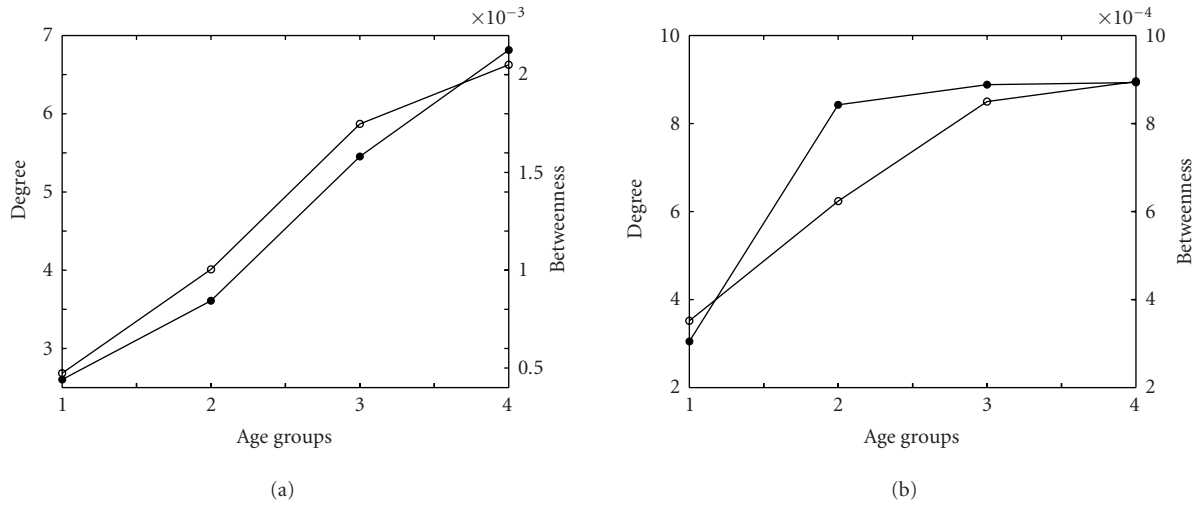


FIGURE 5. Degree and betweenness dependence of protein age. Average degree (left axis, open circle) and average betweenness (right axis, filled circle) of the four age groups of the yeast proteins. Group 1 contains proteins existing only in *S cerevisiae* and hence supposed to be the youngest while group 4 contains proteins existing in the all four branches and hence the oldest [32]. (a) Core data. (b) Full DIP data.

identity of HBLC proteins, but we show that their existence can at least be explained as an unavoidable consequence given certain assumed molecular mechanisms of network growth that involve random link rewiring due to mutations. This, together with the finding that HBLC nodes appear not to be evolutionary older proteins, favors the idea that the presence of HBLC proteins is due to intrinsic, structural, and mechanistic constraints of network growth rather than selective pressure on the growing network. However, to support a contribution of adaptive evolution to this distinct feature of network topology, it will be necessary to obtain larger data sets that can reveal an increased essentiality or higher evolutionary age of HBLC proteins compared with other proteins of the same connectivity class.

The HBLC feature also provides some insight into the modular organization of a large network. Real biological networks have a high clustering coefficient [34], indicating that the immediate neighbors of a given node are likely to be interconnected themselves. As a consequence, there are many alternate paths between two nodes. Betweenness can therefore be relatively small even if a node is highly connected, despite the overall correlation between connectivity and betweenness in the random networks. This could contribute to some variance of betweenness values of a protein with a particular (high) connectivity. On the other hand, the existence of high-betweenness nodes specifically with low connectivity suggests that there are proteins outside such clusters that connect those clusters. Thus, even without a precise definition for what constitutes a particular module, HBLC nodes point to the existence of modularity in the PIN. More specifically, HBLC proteins can be viewed as proteins that link putative network modules within a genome-wide network.

Overall, this work illustrates that nonrandom network topology features represent one of the most simple phenotypic traits, simple enough to stimulate the formulation of generating algorithms, and therefore they provide a useful handle for addressing the fundamental dualism between adaptive evolution and intrinsic constraints in shaping the traits of living organisms.

DATA AND METHODS

Data

Yeast protein pairwise interaction information was from the yeast20040104.lst and ScereCR20040104.tab files, corresponding to the full and core data, respectively, obtained from http://dip.doe-mbi.ucla.edu [15, 16].

Calculation of betweenness centrality B

To calculate B of node i, one first counts the number of shortest paths between two nodes going through node i. Let b_i be the ratio of this number to the total number of shortest paths existing between those two nodes. The sum of b_i over all pairs of nodes in the network gives the betweenness B_i' of the node i. In this paper we use the quantity B_i, the scaled B_i' with respect to the maximum possible B in a network having n nodes, given by

$$B_i = \frac{2B_i'}{(n-1)(n-2)}. \tag{1}$$

B_i is positive and always less than or equal to 1 for any network. Betweenness of the whole graph is defined as the

average of the differences of all B_i from the largest value among the n nodes of the graph.

Model implementation

BA [5], EBA [18], and SV [19, 20] models were implemented as described in the corresponding references. In all these cases we investigated a range of parameters and selected the ones which gave power-law degree distributions. Among them, we searched for the best set of parameters which gave HBLC-type behavior.

Our generative model (DM) was implemented as follows. We start with a few connected nodes, as in [5]. For t number of steps, we apply the link dynamics, the preferential attachment, and the inverse-degree-dependent detachment of links, and then add a node without any links. This process is repeated until the network grows to the desired size. At each step, probability for attachment, p, and detachment, q, are set to be almost equal and adjusted to obtain the desired final mean connectivity. In our simulations we evolved the network till it reached 6000 nodes, corresponding to the approximate total number of genes in *S cerevisiae*. After this evolution process we selected the largest connected component for further network analysis. We selected parameters in such a way that the size of the largest connected component and mean connectivity are similar to that in PIN data. For many sets of parameters, this model produces a scale-free network with HBLC. Figure 2d gives the $k - B$ plot for one such parameter set.

ACKNOWLEDGMENTS

We would like to thank Luis Amaral and Gabriel Eichler for helpful discussions. This work was supported by grants from AFSOR (F49620-01-1-0564) to Sui Huang and from the NIH (CA55833) to Donald E. Ingber.

REFERENCES

[1] Xenarios I, Eisenberg D. Protein interaction databases. *Curr Opin Biotechnol.* 2001;12(4):334–339.

[2] Strogatz SH. Exploring complex networks. *Nature.* 2001;410(6825):268–276.

[3] Erdos P, Renyi A. On random graphs I. *Publicationes Mathematicae.* 1959;6:290–297.

[4] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001;411(6833):41–42.

[5] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science.* 1999;286(5439):509–512.

[6] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science.* 2002;296(5569):910–913.

[7] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science.* 2002;298(5594):824–827.

[8] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science.* 2002;297(5586):1551–1555.

[9] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature.* 1999;402(suppl 6761):C47–C52.

[10] Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. *Nature.* 1998;393(6684):440–442.

[11] Freeman LC. A set of measures of centrality based upon betweenness. *Sociometry.* 1977;40(1):35–41.

[12] Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci USA.* 2002;99(12):7821–7826.

[13] Guimera R, Amaral LAN. Modelling the world-wide airport network. *Eur Phys J B.* 2004;38:381–385.

[14] Goh KI, Oh E, Jeong H, Kahng B, Kim D. Classification of scale-free networks. *Proc Natl Acad Sci USA.* 2002;99(20):12583–12588.

[15] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002;30(1):303–305.

[16] Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics.* 2002;1(5):349–356.

[17] Mewes HW, Frishman D, Guldener U, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 2002;30(1):31–34.

[18] Albert R, Barabasi AL. Topology of evolving networks: local events and universality. *Phys Rev Lett.* 2000;85(24):5234–5237.

[19] Sole RV, Pastor-Satorras R, Smith E, Kepler TB. A model of large-scale proteome evolution. *Adv Compl Syst.* 2002;5(1):43–54.

[20] Vazquez A, Flammini A, Maritan A, Vespignani A. Modeling of protein interaction networks. *Complexus.* 2003;1:38–44.

[21] Berg J, Lässig M, Wagner A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol.* 2004;4:51.

[22] Wagner A. How the global structure of protein interaction networks evolves. *Proc R Soc Lond B Biol Sci.* 2003;270(1514):457–466.

[23] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science.* 2002;296(5568):750–752.

[24] Gould, SJ. *The Structure of Evolutionary Theory.* Cambridge, Mass: Harvard University Press; 2002.

[25] Huang S. Back to the biology in systems biology: what can we learn from biomolecular networks? *Brief Funct Genomic Proteomic.* 2004;2(4):279–297.

[26] Eisenberg E, Levanon EY. Preferential attachment in the protein network evolution. *Phys Rev Lett.* 2003;91(13):138701.
[27] Qin H, Lu HHS, Wu WB, Li WH. Evolution of the yeast protein interaction network. *Proc Natl Acad Sci USA.* 2003;100(22):12820–12824.
[28] Conant GC, Wagner A. Convergent evolution of gene circuits. *Nat Genet.* 2003;34(3):264–266.
[29] Alon U. Biological networks: the tinkerer as an engineer. *Science.* 2003;301(5641):1866–1867.
[30] Wagner A. Does selection mold molecular networks? *Sci STKE.* 2003;2003(202):PE41.
[31] Kauffman S. *The Origins of Order: Self-Organization and Selection in Evolution.* New York, NY: Oxford University Press; 1993.
[32] Wolfram S. *A New Kind of Science.* Champaign, Ill: Wolfram Media; 2002.
[33] Bar-Yam Y. *Dynamics of Complex Systems.* Reading, Mass: Addison-Wesley; 1997.
[34] Wagner A, Fell D. The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci.* 2001;268(1478):1803–1810.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 104–112 • DOI: 10.1155/JBB.2005.104

RESEARCH ARTICLE

Data-Mining Analysis Suggests an Epigenetic Pathogenesis for Type 2 Diabetes

Jonathan D. Wren[1] and Harold R. Garner[2]

[1]*Advanced Center for Genome Technology, Department of Botany and Microbiology, The University of Oklahoma, 101 David L Boren Blvd, Rm 2025, Norman, OK 73019, USA*
[2]*The McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center, 6000 Harry Hines Blvd, Dallas, TX 75390-8591, USA; Departments of Biochemistry and Internal Medicine and Center for Biomedical Inventions, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA*

Received 28 April 2004; revised 3 August 2004; accepted 9 August 2004

The etiological origin of type 2 diabetes mellitus (T2DM) has long been controversial. The body of literature related to T2DM is vast and varied in focus, making a broad epidemiological perspective difficult, if not impossible. A data-mining approach was used to analyze all electronically available scientific literature, over 12 million Medline records, for "objects" such as genes, diseases, phenotypes, and chemical compounds linked to other objects within the T2DM literature but were not themselves within the T2DM literature. The goal of this analysis was to conduct a comprehensive survey to identify novel factors implicated in the pathology of T2DM by statistically evaluating mutually shared associations. Surprisingly, epigenetic factors were among the highest statistical scores in this analysis, strongly implicating epigenetic changes within the body as causal factors in the pathogenesis of T2DM. Further analysis implicates adipocytes as the potential tissue of origin, and cytokines or cytokine-like genes as the dysregulated factor(s) responsible for the T2DM phenotype. The analysis provides a wealth of literature supporting this hypothesis, which—if true—represents an important paradigm shift for researchers studying the pathogenesis of T2DM.

INTRODUCTION

The biomedical literature is vast and growing rapidly, with approximately 12.7 million records in Medline at the time of this writing and growing at a rate of 500 000 new records per year. Time and interest are natural limitations we all share in our awareness of past publications and in keeping up with current ones. As such, our perspective is limited and there is an increasing need for computational methods of literature analysis to gain a broader perspective [1]—particularly ones that lead researchers to reasonable hypotheses that they may not have been able to arrive at independently [2, 3].

Recently, a method was developed for a full-scale analysis of Medline records, with the intention of enabling software to take over some of this tedium so that a global analysis of literature trends could be conducted and statistically relevant associations brought to the attention of the user [4, 5]. This software package, entitled IRIDESCENT (Implicit Relationship IDEntification by in-Silico Construction of an Entity-based Network from Text), "reads" Medline to identify both the primary names (eg, "type 2 diabetes") and synonyms (eg, "adult-onset diabetes") of diseases, genes, phenotypes, and chemical compounds (collectively referred to as "objects") as they appear together within Medline titles and abstracts. As objects appear together more frequently in the same Medline records, their relationship strength is deemed to increase. After processing all of Medline, this database reflects a historical record of research and a summary of what is known or, at least, published. Given an object of interest such as type 2 diabetes mellitus (T2DM), IRIDESCENT first identifies all related objects and then uses these objects to identify implied relationships (Figure 1). That is, it identifies new objects that are unrelated to T2DM yet share many relationships with T2DM. Each of these new objects shares a set of relationships with the original object of interest, but is not itself related. Each set of shared relationships is ranked against a random network model for its statistical relevance. These *implicitly* related objects identified by the system have the potential to offer insight

Correspondence and reprint requests to Jonathan D. Wren, Advanced Center for Genome Technology, Department of Botany and Microbiology, The University of Oklahoma, 101 David L. Boren Blvd., Rm. 2025, Norman, OK 73019, USA, E-mail: jonathan.wren@ou.edu

Table 1. A nonexhaustive list of SNP studies that have reported finding a "significant" association between one or more SNPs and a cohort of T2DM patients. NCBI's sequence viewer was used to obtain flanking sequence. [a] denotes PMID = PubMed ID of paper publishing the SNP. [b] denotes that C allele was also present. [c] means that no dbSNP entry was found in exact map position as given, but the closest dbSNP entries had the same alleles as published, so these were used instead.

Gene	Chr	Region	Mutation	Position	CpG?	PMID[a]	Sequence	dbSNP
IDE	10q23	3′ UTR	multiple	—	—	12765971	—	—
GFPT2	2p13	3′ UTR	multiple	—	—	14764791	—	—
IsI-1	5q11.2	5′ UTR	A→G	−47	—	11978668	—	—
KIR6.2	11p15.1	Coding	E23K	—	—	12643262	—	—
SUR1	11p15.1	Coding	G→A	3819	—	11030411	—	—
PLA2G4A	1q25	Coding	F479L	—	—	12765847	—	—
PTP-1B	20q13.1	Coding	C→T	981	—	11836311	—	—
GFPT2	2p13	Coding	I471V	—	—	14764791	—	—
PPARG2	3p25	Coding	P12A	—	—	12829658	—	—
GLUT2	3q26.1	Coding	A→G	103	—	12017192	—	—
PGC-1	4p15.1	Coding	G482S	—	—	12606537	—	—
NR3C1	5q31	Coding	A→G	1220	—	12864802	—	—
Syntaxin 1A	7q11.23	Coding	T→C	204	—	11719842	—	—
PRKCZ	1p36.33	Intron	G→A	—	—	12970910	—	rs436045
CAPN10	2q37.3	Intron	G→A	—	—	14730479	—	SNP43
UCP2	11q13	Promoter	G→A	−866	Y[b]	12915397	CTGAGGCGT	rs659366
Resistin	19p13.3	Promoter	C→G	−180	Y	12829623	AAGACGGAG	rs1862513
CRP	1q21	Promoter	C→T	−700	Y	12618085	AACACGCGG	SNP133552
PTGS2	1q25.2	Promoter	C→G	−766	Y	12920574	TCCCGCCTC	rs20417
PCK1	20q13.31	Promoter	C→G	−232	Y[c]	14764811	CAACCTTGT	rs6025628
GLUT2	3q26.2	Promoter	A→C	−269	N	12017192	AATCACATG	None
APM1	3q27	Promoter	A→G	−11426	Y[c]	14749263	TCTCGGCTC	rs12631446
APM1	3q27	Promoter	C→G	−11377	N[c]	14749263	ATTACAGGT	rs10937272
TNF-alpha	6p21.3	Promoter	G→A	−308	N[c]	12818408	CATGGGGA	rs1800629
IL-6	7p21	Promoter	G→C	−174	Y	12589429	TTGCCATGC	rs1800795

into novel mechanisms of disease etiology and drug action, among other things. We have applied IRIDESCENT to the study of noninsulin-dependent diabetes mellitus (NIDDM), more commonly known as T2DM, in the hope of elucidating novel relationships pertinent to its pathogenesis.

T2DM is an increasingly prevalent disease in the world, especially the United States, where the number of new patients grew 49% between 1991 and 2000. The economic cost of T2DM is staggering, with the Center for Disease Control estimating costs at $132 billion annually (May 2003 estimates) and affecting more than 6% of the population in the United States (http://www.cdc.gov/diabetes/pubs/estimates.htm). Many factors that correlate with the risk of developing T2DM have been identified, but causality has proven elusive. T2DM has consequently been termed a "complex" disorder [6], thought to be a result of a complex interaction between environmental influence and genetic background. A large number of studies so far have focused upon single-nucleotide polymorphisms (SNPs), the hypothesis being that while their individual associations are weak, they may be acting synergistically and possibly in concert with environmental influences. Table 1, for example, is a nonexhaustive list of 25 different SNPs reported to have "significant" associations with T2DM obtained from a PubMed query on the words "(type 2 diabetes or NIDDM) and polymorphism." The odds any one person would inherit even a significant fraction of these 25, much less than all of them, are extremely low (allele frequencies aside, note their distributions across chromosomes). Thus far, little attention has been paid to the hypothesis that T2DM may be epigenetic in origin, but this is changing as researchers realize that traditional theories on complex disease etiology are lacking [7].

DNA methylation is a fundamentally important phenomenon within eukaryotes, serving as a means to distinguish host DNA from foreign DNA, to determine which strand of DNA is newly replicated [8] and to provide a signal for chromatin condensation such that transcriptional programs can be inactivated, a process especially important during normal development [9]. Loss of methylation in regulatory DNA regions has been an active research area in cancer, with a number of genes known to

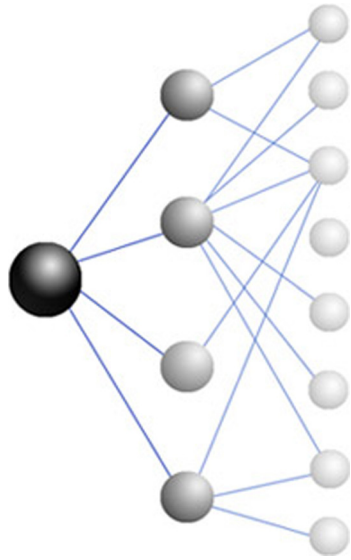


Figure 1. Identifying new relationships that are potentially implicit from known relationships. IRIDESCENT's method of analysis begins with a primary object of interest such as T2DM (black node) and then identifies all co-citations with other objects (gray nodes) observed within Medline. IRIDESCENT then examines all these directly related (gray) nodes in turn for their relationships with other objects (white nodes) that are not themselves related to the primary object. Once all of Medline is analyzed and all implicit relationships (white nodes) are identified, all relationships they share (gray nodes) with the primary node are individually scored for their statistical significance of co-occurrence and collectively scored for statistical significance against a random network model.

be dysregulated from a loss of methylation in certain tumors [10]. Erosion of normal DNA methylation patterns seems to occur in most tissues [11] and is time dependent [12, 13]. Although loss of DNA methylation can be induced chemically (eg, 5-aza-2'-deoxycytidine) or occur as a result of nutrient deficiency (eg, folate), neither of these cases likely apply to T2DM patients, so it is not clear whether other environmental factors could have either a protective or causative effect with regard to T2DM.

MATERIALS AND METHODS

At the time of this analysis, IRIDESCENT was capable of recognizing 33 534 unique objects within text. A total of 2105 of these were cataloged as being directly related to T2DM. IRIDESCENT then analyzed each of these 2105 related objects (schematically illustrated in Figure 1) for their relationships, removing those already in the list of direct relationships. The resulting list contained objects indirectly related to T2DM. That is to say, they shared a large set of relationships with T2DM but were not themselves related to it. At the time of analysis, none of these implicitly related objects were found comentioned with T2DM in the body of any Medline title or abstract. Each implicit relationship was then evaluated by IRIDESCENT based upon the number of relationships it shared with T2DM, relative strength of each relationship, quality of the relationships (statistical probability that each relationship is valid), and the probability that the two objects would share a similar set of relationships by chance, given the relative abundance of both objects and their shared intermediates within the network. A total of 1287 relationships were identified as being shared by the objects "methylation" and "T2DM." Not all of these are necessarily causal, correlative, or even meaningful, but many are. Collectively, they provide evidence that a relationship does exist between epigenetic control and T2DM and enabled us to develop a more comprehensive theory regarding an epigenetic etiology and pathogenesis of T2DM. We will limit the discussion of shared relationships to those we believe are most pertinent (summarized in Figure 2).

RESULTS

IRIDESCENT was used to identify and rank objects within Medline implicitly related to T2DM and identified "methylation" and "chromatin" as top scoring hits (Table 2). Methylation is a fundamentally important phenomenon within eukaryotes for the development and regulation of cellular processes, including the modification and regulation of proteins, lipids, and DNA [14]. Although the relationships linking T2DM and methylation may refer to more than just DNA methylation, much of the discussion here will be focused upon DNA methylation primarily because of the strong link to chromatin and because of the nature of the relationships explored that suggest a permanent change to cellular state is taking place in T2DM as opposed to something caused by a temporary equilibrium shift in molecular methylation capacity.

IRIDESCENT identified a number of common phenotypes in the onset and pathology of T2DM that are also shared by diseases associated with a change in methylation state. These shared relationships offer a perspective on some of the puzzling properties of T2DM not easily explained by environmental or genetic mutation models. For example, the onset of T2DM varies, but usually occurs later in life and the probability of affliction generally increases with age. This pattern is also characteristic of epigenetic disorders such as aberrant expression of X-linked genes [15], onset of Huntington's disease [16], and the oncogenesis of tumors [17, 18]. Not all late-onset illnesses are caused by epigenetic changes, but most others share physical accumulations that are unique to the disease, such as the accumulation of amyloid precursor proteins in Alzheimer's disease [19] or Lewy bodies in Parkinsons [20]. T2DM is highly correlated with the presence of obesity and advanced glycosylation end products (AGEs), but neither is a requirement for its development nor unique to it as a disease. T2DM also varies in its severity, generally increasing over time. This is a phenotype shared with some tumors that have undergone methylation changes

Table 2. Objects linked to T2DM solely by virtue of relationships they share within Medline. At the time of analysis, none of these objects were documented in Medline to have a relationship with T2DM (shown at top as a positive control for the query). The nature of each of these implicit relationships varies and must be determined by examination of the intermediate connections. The expected value represents how many shared relationships would be expected given a randomly connected network of relationships with the same properties of the literature-derived one. The quality score represents a statistical weighting of co-mention frequency to reflect confidence that the relationship is not of a trivial nature. The observed to expected ratio (Obs/exp) provides a ranking of how statistically exceptional any given set of shared relationships is.

Rank	Shared relationships	Implicit relationship	Quality	Expect	Obs/exp
—	2105	T2DM	1421	329	4.32
1	1361	Endotoxin	1054	308	3.42
2	1312	Hydrocortisone	991	296	3.35
3	1301	Neuroblastoma	975	339	2.88
4	1287	Methylation	959	346	2.77
5	1256	Chromatin	938	339	2.77

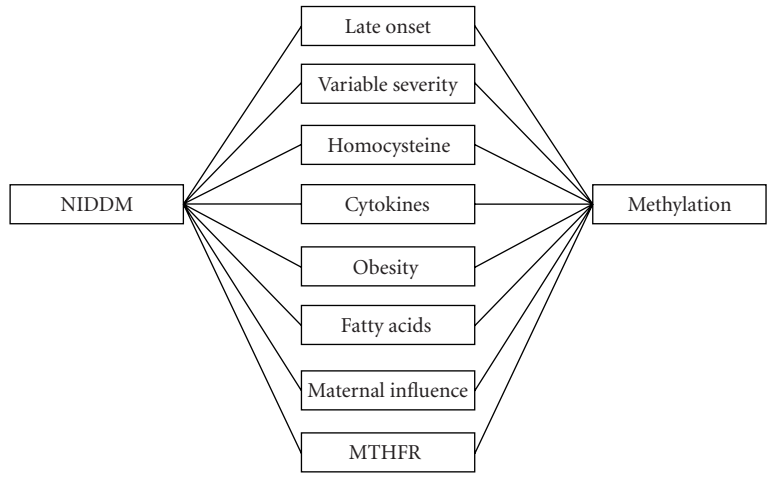


Figure 2. Important relationships shared by methylation and T2DM (referred to as "NIDDM" in the system). A total of 1287 co-cited objects were identified between the two, about 959 of these reflect actual relationships of a nontrivial nature. Only relationships emphasized within this report are shown here. A full list is available online at http://innovation.swmed.edu/IRIDESCENT/NIDDM_theory.htm.

in promoter sequences, leading to higher gene expression and a more aggressive phenotype [17]. Another interesting observation about T2DM is the "maternal effect" in which T2DM patients report a higher frequency of maternal history of diabetes [21]. While this is not without controversy [22], such an effect could be explained if de novo methylation of DNA sequences during development was due to maternal influence. This type of phenomenon, in fact, has been observed in mice [23, 24].

METABOLIC CHANGES IN T2DM

IRIDESCENT also identified a number of metabolic alterations in the body's ability to methylate DNA that correlate with the existence of or predisposition to T2DM. For example, elevated levels of homocysteine have been found in T2DM patients, correlating with increased severity of the disease as defined by mortality [25]. Homocysteine is a critical metabolic intermediate responsible for carrying out methylation reactions, and the elevated serum levels of it are also correlated with DNA hypomethylation [26]. It has also been reported that sulfur-poor diets that force synthesis of cysteine from methionine predispose individuals to type 2 diabetes later in life [27, 28]. Since methionine affects S-adenosyl methionine (SAM), which is the methyl donor for the methylation of newly synthesized DNA, these individuals develop with an impaired ability to establish de novo DNA methylation patterns. Genetic factors that lead to deficiencies in the methylation pathway have also been shown to predispose individuals to develop T2DM. There is a well-known polymorphism (C677T) in the methylenetetrahydrofolate reductase (MTHFR) gene that decreases its efficiency, leading to a global hypomethylation of DNA [29]. Individuals with this mutation are also predisposed to develop T2DM and other complications of the metabolic syndrome [30]. It has also been found that lowering the amount of dietary methyl-providing compounds

provided to mice during pregnancy leads to a global hypomethylation of DNA in their offspring and a corresponding modification of offspring hair coat color (increased yellow versus agouti). Expression of this yellow coat color is associated with increased risks of obesity, diabetes, and cancer [24]. And finally, it has also been shown that aberrant methylation patterns have been shown to induce diabetic symptoms in another form of diabetes, transient neonatal diabetes mellitus (TNDM), which is a result of genetic imprinting [31]. The same imprinted region responsible for TNDM, however, is not known to be responsible for T2DM [32].

A number of these metabolic correlations were very recently noted independently of our own analysis by a company proposing to do a genome-wide scan for alterations in DNA methylation patterns [33], giving further credence to the idea that changes in DNA methylation is a causative factor in the etiology of T2DM. While identifying an etiology is important, even in general terms, perhaps equally or more important is elucidating the causative factors in the pathogenesis of T2DM. If epigenetic alterations are responsible for T2DM, then at least three questions naturally arise: first, what secreted factors are responsible for the T2DM phenotype; second, what tissue type(s) is responsible for expressing the factors that induce the T2DM phenotype; and third, what environmental factors could lead to a loss of methylation and consequent dysregulation of the secreted factors.

CAUSAL FACTOR ANALYSIS

A possible answer for the first question above comes from the highest scoring object on IRIDESCENT's list of implicitly related objects (Table 2): endotoxins. While endotoxins are not known to be associated or causal in T2DM, they have been shown to induce obesity and insulin resistance [34]. Most of the relationships shared between T2DM and endotoxins are objects that either affect or are involved in the immune response, especially cytokines and inflammatory factors. Expression of acute-phase markers such as C-reactive protein (CRP), and proinflammatory cytokines such as IL-6 and TNF-alpha is highly correlated with the presence and severity of T2DM symptoms [35, 36]. These proinflammatory cytokines are also positively correlated with obesity [37]. Furthermore, TNF-alpha has been found to induce insulin resistance [38]. Indeed, there is a growing body of evidence that cytokines, more specifically the proinflammatory cytokines, could be responsible for the T2DM phenotype. It has been observed, for example, that a reversal of T2DM symptoms can be induced by disruption of the inflammatory pathway with high doses of aspirin [39]. Troglitazone, a widely used medication to treat T2DM, has also been found to have anti-inflammatory properties [40], and the lifestyle changes of exercise and dietary changes prescribed to T2DM patients that have been successful in reversing T2DM phenotypes have also been associated with decreases in inflammatory cytokines [41, 42].

If a gene(s) were to be dysregulated, it would likely happen via demethylation of its (or their) promoters. It is possible that some of the SNP association studies may have been linked to promoters with fewer methylatable sites. Within the 25 SNP studies that were surveyed earlier, 10 were found within the promoter region. None of them, however, made the association between the reported SNP and the possibility that the SNP would alter the number of CpGs in the promoter. So to examine this possibility, we used NCBI's sequence viewer to obtain the surrounding sequence for these SNPs. We found that, of the 10 promoter SNPs, 7 altered the number of CpGs present within the promoter (see Table 1). Of these 7, 3 were proinflammatory cytokines associated with T2DM-resistin [43], CRP [44], and IL-6 [45]. Interestingly, IL-6 controls both resistin [46] and CRP expression, with the -174 SNP in IL-6 having been linked to heritably high CRP levels [47]. No study has been done yet to see if this IL-6 polymorphism can be linked to resistin levels as well.

TISSUE OF ORIGIN ANALYSIS

Whereas proinflammatory cytokines have been implicated as a causal factor in T2DM, it is known that besides B cells and T cells, adipocytes and endothelial cells are the only other cell types known to normally produce cytokines. We see that within T cells, cytokine expression is determined by DNA methylation patterns [48] and can be altered by demethylating agents [49]. Neither T cells nor B cells seem a likely candidate since they are not very metabolically active in their naïve or memory forms, and their more active differentiated forms are relatively short lived. Adipocytes, however, are the primary repository for lipids and produce cytokines in proportion to factors such as their size and surrounding obesity [50]. Interestingly, a study by Benjamin and Jost demonstrated that short-chain fatty acids (SCFAs) can promote the demethylation of actively transcribed regions [51]. SCFAs can also affect chromatin structure by inhibiting HDAC, causing hyperacetylation of histones [52] and making regions of DNA more accessible to transcription factors. SCFAs are not normally present in high concentrations within adipocytes, but are normal metabolic byproducts of the long-chain fatty acids stored within. Since the rate of lipolysis within adipocytes is increased in T2DM [53], and can be induced by factors such as TNF-alpha already known to be elevated in T2DM [54], this would have an effect upon the relative concentrations of SCFAs within adipocytes. Higher amounts of SCFA metabolites within adipocytes might provide an environment in which loss of DNA methylation could occur and, coupled with active transcriptional activity, could lead to the hypomethylation and consequent dysregulation of cytokines or cytokine-like factors that lead to T2DM. We see suggestive evidence of this in a study by Laimer et al involving IL-6 and TNF-alpha levels in 20 women before and 1 year after gastric banding surgery. They found that the levels

of other obesity markers such as CRP declined, while IL-6 and TNF-alpha did not [55].

Within the proposed model, the etiology of T2DM occurs within adipocytes, involving a gradual loss of DNA methylation around the promoters of cytokines and/or cytokine-like factors normally secreted by the adipocyte. This loss of methylation is favored under the conditions provided by obesity and is caused by transcriptional activity. The subsequent loss of methylation leads to dysregulation of these factors, resulting in a constitutive increase in the production of cytokines from adipocytes. Negative regulatory factors decrease the expression of these factors, enabling management of the T2DM phenotype, but only as long as they are present.

ETIOLOGICAL MODELS OF T2DM

We examine this new proposed model in the context of the three dominant models for the etiology and pathogenesis of T2DM: genetic, environmental, and a complex interaction of both factors. Genetic studies have shown that inheritance plays a role in determining an individual's risk of developing T2DM [56]. Linkage studies, while delineating a number of potential susceptibility regions, have yet to be successful in identifying a specific gene or set of genes responsible for the most popular form of T2DM, despite the large cohorts involved. The well-established correlation between obesity and T2DM also indicates that environmental variables affect the pathogenesis of T2DM. The prevailing "complex disease" theory is that the onset of T2DM is caused by one or more environmental variables acting upon a genetic background, of which there may be many contributing genes [6]. This theory explains how susceptibility to T2DM correlates with genetic background, such as race, as well as with environmental variables such as diet and exercise. But there are at least two observations about the nature of T2DM that the complex disease model does not explain while the epigenetic model does: time dependency and systemic memory.

Even when environmental variables are present on a susceptible genetic background, the onset of T2DM is still time dependent. That is to say, the risk of developing T2DM is positively correlated with age. The complex disease model does not easily explain this except to postulate an as-yet-unknown "trigger" event, such as an infection. Even if this were true, it would not explain the persistence of T2DM after onset. T2DM is diagnozed by the levels of insulin resistance and glucose intolerance experienced by a patient, levels which can be altered to prediabetic levels by sufficient changes in lifestyle. T2DM, however, cannot be reversed [57]. None of the existing models account for a mechanism by which the body can "remember" its state. The methylation status of genes, however, is intended to be a relatively persistent phenomenon, responsible for committing cells into their differentiated states [58]. Given that loss of DNA methylation is correlated with age [59], that the number of methylated sites in a genome is determined by inheritance, and that loss of methylation can be affected by environmental variables, the proposed epigenetic model merits serious consideration. An excellent review was recently published, contrasting the properties of epigenetic disorders with the other models of disease etiology discussed here [60].

DISCUSSION

Contrary to the mutation-centric model, which assumes alterations in function or activity based upon either somatic or inherited mutations in DNA, an epigenetic model implies dysregulation of a gene or set of genes. Thus, phenotypes resulting from the expression of such genes would make biological sense under other physiological conditions. Preventing energy influx into cells by inducing insulin resistance makes sense when considered within the context of the role of the immune system. Acquired immunity in the form of B-cell maturation and antibody production takes time during which pathogens are able to replicate. Part of the early immune response consists of an increase in the presence of proinflammatory cytokines within the circulating bloodstream [61, 62]. It would make sense that one role of these early responders would be to stem the influx of resources like glucose into cells to prevent their utilization by invading pathogens. Since adipocytes contain a large reservoir of energy, this makes them ideal targets for invading pathogens and could necessitate their taking a more active role in fighting infection beyond that of other somatic cells.

Identifying expression changes via microarray analysis and subsequently examining the methylation status of their promoters can obtain a candidate list for genes that have undergone epigenetic dysregulation. If this theory is ultimately shown to be correct, it will allow us the ability to diagnoze the current level of epigenetic progression towards T2DM in patients and offer hope for a T2DM cure that could not be easily provided in a mutation-centric model. It is not apparent how region-specific methylation could be reintroduced to affected regions, but since de novo methylation is a normal process during development and certain viruses can "shut off" the expression of immune-related genes by hypermethylation, it stands to reason that the mechanism to do so is already in place.

ACKNOWLEDGMENTS

We would like to thank Dr Roger Unger for his very helpful review of the manuscript and suggestions. This work was funded in part by NSF-EPSCoR Grant no EPS-0132534, NIH/NCI Grant no R33 CA81656, and NIH/NHLBI Grant no P50 CA70907.

REFERENCES

[1] Yandell MD, Majoros WH. Genomics and natural language processing. *Nat Rev Genet.* 2002;3(8):601–610.

[2] Bray D. Reasoning for results. *Nature.* 2001; 412(6850):863.

[3] Blagosklonny MV, Pardee AB. Conceptual biology: unearthing the gems. *Nature.* 2002;416(6879):373.

[4] Wren JD, Garner HR. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics.* 2004;20(2):191–198.

[5] Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics.* 2004;20(3):389–398.

[6] Marx J. Unraveling the causes of diabetes. *Science.* 2002;296(5568):686–689.

[7] Dennis C. Epigenetics and disease: altered states. *Nature.* 2003;421(6924):686–688.

[8] Woodcock DM, Simmons DL, Crowther PJ, Cooper IA, Trainor KJ, Morley AA. Delayed DNA methylation is an integral feature of DNA replication in mammalian cells. *Exp Cell Res.* 1986;166(1):103–112.

[9] Attwood JT, Yung RL, Richardson BC. DNA methylation and the regulation of gene transcription. *Cell Mol Life Sci.* 2002;59(2):241–257.

[10] Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumors. *J Pathol.* 2002;196(1):1–7.

[11] Cooney CA. Are somatic cells inherently deficient in methylation metabolism? A proposed mechanism for DNA methylation loss, senescence and aging. *Growth Dev Aging.* 1993;57(4):261–273.

[12] Wareham KA, Lyon MF, Glenister PH, Williams ED. Age related reactivation of an X-linked gene. *Nature.* 1987;327(6124):725–727.

[13] Fuke C, Shimabukuro M, Petronis A, et al. Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: an HPLC-based study [published correction appears in *Ann Hum Genet*. 2005;69(pt 1):134]. *Ann Hum Genet.* 2004;68(pt 3):196–204.

[14] Hoffman RM. Methioninase: a therapeutic for diseases related to altered methionine metabolism and transmethylation: cancer, heart disease, obesity, aging, and Parkinson's disease. *Hum Cell.* 1997;10(1):69–80.

[15] Anderson CL, Brown BJ. Variability of X chromosome inactivation: effect on levels of TIMP1 RNA and role of DNA methylation. *Hum Genet.* 2002;110(3):271–278.

[16] Reik W, Maher ER, Morrison PJ, Harding AE, Simpson SA. Age at onset in Huntington's disease and methylation at D4S95. *J Med Genet.* 1993;30(3):185–188.

[17] Kim YI, Giuliano A, Hatch KD, et al. Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer.* 1994;74(3):893–899.

[18] Qu G, Dubeau L, Narayan A, Yu MC, Ehrlich M. Satellite DNA hypomethylation vs overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. *Mutat Res.* 1999;423(1–2):91–101.

[19] Hardy J. The Alzheimer family of diseases: many etiologies, one pathogenesis? *Proc Natl Acad Sci USA.* 1997;94(6):2095–2097.

[20] Nussbaum RL, Polymeropoulos MH. Genetics of Parkinson's disease. *Hum Mol Genet.* 1997;6(10): 1687–1691.

[21] Alcolado JC, Laji K, Gill-Randall R. Maternal transmission of diabetes. *Diabet Med.* 2002;19(2):89–98.

[22] Thorand B, Liese AD, Metzger MH, Reitmeir P, Schneider A, Lowel H. Can inaccuracy of reported parental history of diabetes explain the maternal transmission hypothesis for diabetes? *Int J Epidemiol.* 2001;30(5):1084–1089.

[23] Pickard B, Dean W, Engemann S, et al. Epigenetic targeting in the mouse zygote marks DNA for later methylation: a mechanism for maternal effects in development. *Mech Dev.* 2001;103(1–2):35–47.

[24] Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB J.* 1998;12(11):949–957.

[25] Stehouwer CD, Gall MA, Hougaard P, Jakobs C, Parving HH. Plasma homocysteine concentration predicts mortality in non-insulin-dependent diabetic patients with and without albuminuria. *Kidney Int.* 1999;55(1):308–314.

[26] Yi P, Melnyk S, Pogribna M, Pogribny IP, Hine RJ, James SJ. Increase in plasma homocysteine associated with parallel increases in plasma S-adenosylhomocysteine and lymphocyte DNA hypomethylation. *J Biol Chem.* 2000;275(38):29318–29323.

[27] Rees WD. Manipulating the sulfur amino acid content of the early diet and its implications for long-term health. *Proc Nutr Soc.* 2002;61(1):71–77.

[28] Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet.* 1986;1(8489):1077–1081.

[29] Stern LL, Mason JB, Selhub J, Choi SW. Genomic DNA hypomethylation, a characteristic of most cancers, is present in peripheral leukocytes of individuals who are homozygous for the C677T polymorphism in the methylenetetrahydrofolate reductase gene. *Cancer Epidemiol Biomarkers Prev.* 2000;9(8):849–853.

[30] Benes P, Kankova K, Muzik J, et al. Methylenetetrahydrofolate reductase polymorphism, type II diabetes mellitus, coronary artery disease, and essential hypertension in the Czech population. *Mol Genet Metab.* 2001;73(2):188–195.

[31] Temple IK, Gardner RJ, Robinson DO, et al. Further evidence for an imprinted gene for neonatal diabetes localised to chromosome 6q22-q23. *Hum Mol Genet.* 1996;5(8):1117–1121.

[32] Shield J, Owen K, Robinson DO, et al. Maturity onset diabetes of the young (MODY) and early onset type II diabetes are not caused by loss of imprinting at the transient neonatal diabetes (TNDM) locus. *Diabetologia.* 2001;44(7):924.

[33] Maier S, Olek A. Diabetes: a candidate disease for efficient DNA methylation profiling. *J Nutr.* 2002;132(suppl 8):2440S–2443S.

[34] Nilsson C, Larsson BM, Jennische E, et al. Maternal endotoxemia results in obesity and insulin resistance in adult male offspring. *Endocrinology.* 2001;142(6):2622–2630.

[35] Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA.* 2001;286(3):327–334.

[36] Kern PA, Ranganathan S, Li C, Wood L, Ranganathan G. Adipose tissue tumor necrosis factor and interleukin-6 expression in human obesity and insulin resistance. *Am J Physiol Endocrinol Metab.* 2001;280(5):E745–751.

[37] Das UN. Is obesity an inflammatory condition? *Nutrition.* 2001;17(11–12):953–966.

[38] Moller DE. Potential role of TNF-alpha in the pathogenesis of insulin resistance and type 2 diabetes. *Trends Endocrinol Metab.* 2000;11(6):212–217.

[39] Yuan M, Konstantopoulos N, Lee J, et al. Reversal of obesity- and diet-induced insulin resistance with salicylates or targeted disruption of Ikkbeta [published correction appears in *Science.* 2002;295(5553):277]. *Science.* 2001;293(5535):1673–1677.

[40] Aljada A, Garg R, Ghanim H, et al. Nuclear factor-kappaB suppressive and inhibitor-kappaB stimulatory effects of troglitazone in obese patients with type 2 diabetes: evidence of an antiinflammatory action? *J Clin Endocrinol Metab.* 2001;86(7):3250–3256.

[41] Ziccardi P, Nappo F, Giugliano G, et al. Reduction of inflammatory cytokine concentrations and improvement of endothelial functions in obese women after weight loss over one year. *Circulation.* 2002;105(7):804–809.

[42] Pedersen BK, Steensberg A, Fischer C, Keller C, Ostrowski K, Schjerling P. Exercise and cytokines with particular focus on muscle-derived IL-6. *Exerc Immunol Rev.* 2001;7:18–31.

[43] Smith SR, Bai F, Charbonneau C, Janderova L, Argyropoulos G. A promoter genotype and oxidative stress potentially link resistin to human insulin resistance. *Diabetes.* 2003;52(7):1611–1618.

[44] Wolford JK, Gruber JD, Ossowski VM, et al. A C-reactive protein promoter polymorphism is associated with type 2 diabetes mellitus in Pima Indians [published correction appears in *Mol Genet Metab.* 2003;79(3):231]. *Mol Genet Metab.* 2003;78(2):136–144.

[45] Vozarova B, Fernandez-Real JM, Knowler WC, et al. The interleukin-6 (-174) G/C promoter polymorphism is associated with type-2 diabetes mellitus in Native Americans and Aaucasians. *Hum Genet.* 2003;112(4):409–413.

[46] Kaser S, Kaser A, Sandhofer A, Ebenbichler CF, Tilg H, Patsch JR. Resistin messenger-RNA expression is increased by proinflammatory cytokines in vitro. *Biochem Biophys Res Commun.* 2003;309(2):286–290.

[47] Vickers MA, Green FR, Terry C, et al. Genotype at a promoter polymorphism of the interleukin-6 gene is associated with baseline levels of plasma C-reactive protein. *Cardiovasc Res.* 2002;53(4):1029–1034.

[48] Wilson CB, Makar KW, Perez-Melgosa M. Epigenetic regulation of T cell fate and function. *J Infect Dis.* 2002;185(suppl 1):S37–45.

[49] Rothenburg S, Koch-Nolte F, Thiele HG, Haag F. DNA methylation contributes to tissue- and allele-specific expression of the T-cell differentiation marker RT6. *Immunogenetics.* 2001;52(3–4):231–241.

[50] Coppack SW. Pro-inflammatory cytokines and adipose tissue. *Proc Nutr Soc.* 2001;60(3):349–356.

[51] Benjamin D, Jost JP. Reversal of methylation-mediated repression with short-chain fatty acids: evidence for an additional mechanism to histone deacetylation. *Nucleic Acids Res.* 2001;29(17):3603–3610.

[52] Sealy L, Chalkley R. The effect of sodium butyrate on histone modification. *Cell.* 1978;14(1):115–121.

[53] Foley JE, Kashiwagi A, Verso MA, Reaven G, Andrews J. Improvement in in vitro insulin action after one month of insulin therapy in obese noninsulin-dependent diabetics. Measurements of glucose transport and metabolism, insulin binding, and lipolysis in isolated adipocytes. *J Clin Invest.* 1983;72(6):1901–1909.

[54] Green A, Dobias SB, Walters DJ, Brasier AR. Tumor necrosis factor increases the rate of lipolysis in primary cultures of adipocytes without altering levels of hormone-sensitive lipase. *Endocrinology.* 1994;134(6):2581–2588.

[55] Laimer M, Ebenbichler CF, Kaser S, et al. Markers of chronic inflammation and obesity: a prospective study on the reversibility of this association in middle-aged women undergoing weight loss by surgical intervention. *Int J Obes Relat Metab Disord.* 2002;26(5):659–662.

[56] Haffner SM, Stern MP, Hazuda HP, Pugh JA, Patterson JK. Hyperinsulinemia in a population at high risk for non-insulin-dependent diabetes mellitus. *N Engl J Med.* 1986;315(4):220–224.

[57] Nathan DM. Prevention of long-term complications of non-insulin-dependent diabetes mellitus. *Clin Invest Med.* 1995;18(4):332–339.

[58] Michalowsky LA, Jones PA. DNA methylation and differentiation. *Environ Health Perspect.* 1989;80: 189–197.
[59] Catania J, Fairweather DS. DNA methylation and cellular ageing. *Mutat Res.* 1991;256(2–6):283–293.
[60] Bjornsson HT, Fallin MD, Feinberg AP. An integrated epigenetic and genetic approach to common human disease. *Trends Genet.* 2004;20(8):350–358.
[61] Toossi Z. The inflammatory response in mycobacterium tuberculosis infection. *Arch Immunol Ther Exp (Warsz).* 2000;48(6):513–519.
[62] Cooper MA, Fehniger TA, Ponnappan A, Mehta V, Wewers MD, Caligiuri MA. Interleukin-1beta costimulates interferon-gamma production by human natural killer cells. *Eur J Immunol.* 2001;31(3):792–801.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 113–123 • DOI: 10.1155/JBB.2005.113

RESEARCH ARTICLE

Combining Information From Multiple Data Sources to Create Multivariable Risk Models: Illustration and Preliminary Assessment of a New Method

Greg Samsa,[1,2] Guizhou Hu,[3] and Martin Root[3]

[1]*Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, USA*
[2]*Center for Clinical Health Policy Research, Duke University Medical Center, Durham, NC 27705, USA*
[3]*BioSignia, Inc, 1822 East NC Highway 54, Durham, NC 27713, USA*

Received 5 February 2004; revised 29 March 2004; accepted 6 April 2004

A common practice of metanalysis is combining the results of numerous studies on the effects of a risk factor on a disease outcome. If several of these composite relative risks are estimated from the medical literature for a specific disease, they cannot be combined in a multivariate risk model, as is often done in individual studies, because methods are not available to overcome the issues of risk factor colinearity and heterogeneity of the different cohorts. We propose a solution to these problems for general linear regression of continuous outcomes using a simple example of combining two independent variables from two sources in estimating a joint outcome. We demonstrate that when explicitly modifying the underlying data characteristics (correlation coefficients, standard deviations, and univariate betas) over a wide range, the predicted outcomes remain reasonable estimates of empirically derived outcomes (gold standard). This method shows the most promise in situations where the primary interest is in generating predicted values as when identifying a high-risk group of individuals. The resulting partial regression coefficients are less robust than the predicted values.

INTRODUCTION

We propose essentially a multivariate metanalytic technique. Many diseases have numerous risk factors, which are often studied in diverse cohorts with only a limited number of risk factors in each. We here propose a method of combining univariate relative risks (betas) from diverse studies into multivariate models.

Metanalysis has proven to be a powerful tool, when handled appropriately, to summarize previous medical research on a common topic, including epidemiologic research [1, 2, 3]. Several issues need to be carefully considered in reaching conclusions from the metanalysis of epidemiologic studies. Studies are often heterogeneous in their findings [4], which can even be considered a benefit in understanding the source of differences in research findings [5]. Publication bias must also be evaluated in a field in which the decision to publish, the quality and size of the study, and the publishing journal's reputation are strongly interconnected [6].

All authorities agree that the best means of combining effect estimates is by a pooled analysis where the separate study datasets are combined together with possible confounders [3], especially if this pooling is planned prospectively. In general though, effect estimates (β coefficients) are combined from published reports. Univariate betas are combined as in the example of Ernst et al, who considered fibrinogen as a risk factor for cardiovascular disease using univariate and age-adjusted parameters [7]. More commonly, multivariate-adjusted odds ratios and relative risks are used as by Etminan et al on the effects of NSAIDs on Alzheimer's disease onset [8] by Vincent et al on hypoalbuminemia in acute illness [9], or by Danesh et al in summarizing various plasma risk factors and heart disease [10].

Our method of preparing multivariate risk models by metanalysis suggests a comparison with multivariate metanalysis. Unfortunately this term covers several techniques, none of which are similar to ours. In some cases it refers to a metanalysis that considers several similar outcomes with the same risk factor [11, 12, 13]. Another technique, also called metaregression, is essentially a weighted multivariate analysis of all the confounders (and possible sources of heterogeneity) in the summarized studies [11].

Correspondence and reprint requests to Martin Root, BioSignia, Inc, 1822 East NC Highway 54, Durham, NC 27713, USA, E-mail: mroot@biosignia.com

Most true multivariate metanalyses are pooled analyses of multiple studies together with possible confounders. Farrer et al used a pooled analysis to estimate the effect of the interaction between age, sex, and ethnicity on the effect of apolipoprotein *E*4 as a predictor of Alzheimer's disease [14].

There would be clear benefit to a metanalytic technique that could combine univariate risk factors for a disease obtained from different studies. In the metanalysis of Danesh et al, four blood parameters were determined to be significantly correlated with heart disease risk [10]. However, it was impossible to combine those in any meaningful way to determine their joint predictive power or to determine their independence from one another. Oftentimes a researcher simply wants to add a single risk factor to an established multivariate risk model. In a recent example, coronary artery calcium score was combined with the Framingham score for predicting heart disease [15]. Previous studies had showed that both scores were strongly predictive of heart disease but to combine them took a 7-year study with 1461 subjects on whom both scores were collected.

In another example, Gail et al developed a model for the prediction of breast cancer onset [16]. This was subsequently modified by statisticians of the National Surgical Adjuvant Breast and Bowel Project (NSABP) to define eligibility criteria for the Breast Cancer Prevention Trial [17]. They used a new source of data, the Surveillance, Epidemiology, and End Results (SEER) Program for new incidence rates for invasive breast cancer to replace the original incidence rates for total breast cancer. They also used the same new dataset to determine race-specific incidence rates to replace the Whites-only rate from the original model [18]. Their method is reported in an NSABP document [19]. In a final example, the need was simply to modify the Framingham score [20] for heart disease to a new, lower-risk, population. Based on the assumption that the multivariate betas are similar across populations, some authors have suggested changing the equation intercept to reflect the underlying incidence rate of the new population [21, 22]. Others have suggested also modifying the risk factor values themselves by using the prevalence rates of the new population [23, 24]. All of these examples involve combining evidence from different sources into unique multivariate risk models based broadly on the assumption that the correlations among risk factors and between risk factors and the endpoint were not significantly different between data sources or populations.

Matchar et al used a variety of techniques and datasets to develop the Stroke Prevention Policy Model (SPPM) [25]. While they concede difficulties and shortcomings of such an approach, they conclude, for clinical and economic applications, "that despite the difficulties in developing comprehensive models, ..., the benefits of such models exceed the costs of continuing to rely on more conventional methods." The SPPM was then used in demonstrating the economic benefit of a stroke treatment's short-term effect on long-term economic outcomes [26].

The method we are about to introduce can also be used in datasets with a large fraction of missing values. Generally two strategies exist for such situations, either using modeling techniques to extract data from observations with incomplete data or data imputation [27]. Zhao et al have introduced a joint estimating equation that robustly estimates effect size in multivariate models with missing data [28]. Steyerberg et al address the related problem of underpowered small studies [29]. They describe a method to combine results from the medical literature with results from individual patient data and conclude "that prognostic models {from small studies} may benefit substantially from explicit incorporation of literature data."

We have developed a new statistical method to address the question of combining estimates of partial regression parameters across datasets. This method is intended to provide an approximate solution in the circumstances illustrated above. The performance of this method is assessed via simulation.

METHODS

Notation

The continuous outcome variable is denoted by Y, and its predicted value by $\hat{Y}$. We first consider a "gold-standard" dataset including all the predictors of interest. Information from the gold-standard dataset is denoted with an asterisk. In practice this gold-standard dataset will not be available, and the predictors of interest will be distributed across multiple "candidate" datasets. The goal will be to estimate, using information from the candidate datasets, the regression relationship between the risk factors and the outcome that would have been observed if the complete dataset had been available.

Denote the vector of predictors in the gold-standard dataset by

$$X^* = (X_0^*, X_1^*, \ldots, X_Q^*) \tag{1}$$

with the first element $X_0^* = 1$ being included in order to estimate the intercept and the remaining Q predictors being of primary interest. The multivariable regression of Y^* on X^* is

$$\hat{Y}^* = \hat{\beta}^* X^*, \tag{2}$$

where $\hat{Y}^*$ is the predicted value of Y and β^* is estimated in the usual way as

$$B^* = (X^{*'} X^*)^{-1} (X^{*'} Y^*). \tag{3}$$

In other words, the multivariable regression equation observed in the data is

$$\hat{Y}^* = a^* + b_1^* X_1^* + b_2^* X_2^* + \cdots + b_Q^* X_Q^*, \tag{4}$$

where, for example, b_1^* estimates β_1^*, and so forth. We will focus on the regression coefficients observed in the data—that is, on B^* rather than β^*.

In the above equations, the estimates of the partial regression coefficients produced from the gold-standard dataset are

$$B^* = (b_1^*, b_2^*, \ldots, b_Q^*). \quad (5)$$

In contrast, each of the "univariable" regression coefficients, denoted, for example, by b_{u1}^*, is the result of fitting a univariable regression model with a single predictor—for example,

$$\hat{Y}^* = a_{1u}^* + b_{1u}^* X_1^*. \quad (6)$$

We assume that there are Q univariable regression coefficients available for use, one from each candidate dataset. The vector of univariable regression coefficients from the candidate datasets is denoted as

$$B_u = (b_{u1}, b_{u2}, \ldots, b_{uQ}). \quad (7)$$

In practice, the observed values of B_u and B_u^* can differ because of (a) differences between β_u and β_u^* and (b) sampling variability within each of the datasets in question.

For concreteness, our goal is to estimate the multivariable regression model, as summarized through the set of Q partial regression coefficients $(b_1^*, b_2^*, \ldots, b_Q^*)$ and the predicted values $\hat{Y}^*$ that could have been produced were the gold-standard dataset is available. In the absence of this gold-standard dataset, we assume that from Q candidate datasets, each containing exactly one predictor variable, we have available its standard deviation (for study j, denoted by s_j and combined into a vector S), as well as its univariable regression coefficient (for study j, denoted by b_{uj}, and combined into a vector B_u). We also assume that from one or more additional datasets, the various first-order correlations between each set of predictors (denoted by r_{ij}, and combined into a matrix R) are available. (These additional datasets need not contain Y.)

It is important to note that this formulation of the problem includes, as a special case, the situation where the various studies include overlapping risk factors. In particular, for each study (whose number need not equal Q) we could estimate a set of univariable regression coefficients—that is, one coefficient per risk factor per study. The additional problem induced by overlapping predictors is that different estimates of b_{uj} will be available for some or all of the risk factors, and that each of these estimates must somehow be reconciled into a single "best" estimate. In this case, we might (1) use standard metanalytic techniques to combine the various estimates of b_{uj} or (2) select the b_{uj} from the "best" available datasets. Estimates of S and R that reconcile multiple estimates can be generated in a similar fashion.

Proposed approach

To illustrate our proposed approach, termed the univariable synthesis method, first consider the gold-standard dataset. Within this dataset, we can calculate (1) univariable regression coefficients for each predictor, denoted by B_u^*, (2) standard deviations for each predictor, denoted by S^*, and (3) the set of all pairwise correlations between the predictor variables, denoted by R^*. Denoting element-wise multiplication by "$\cdot$," and element-wise division by "/," the core of the univariable synthesis method relies on noting that $(b_1^*, b_2^*, \ldots, b_Q^*)$—that is, the portion of B^* excluding the intercept—can also be estimated by [30, equation 1]

$$B^* = \frac{(R^{*-1}(B_u^* \cdot S^*))}{S^*}. \quad (8)$$

The basic idea behind the univariable synthesis method is that, when candidate datasets must be used, the various elements of B_u, R, and S can nevertheless be accumulated across these multiple data sources. In order to do so, it must be assumed that the relevant standard deviations, univariable regression coefficients, and correlations are comparable across studies. (More precisely, we are assuming, in analogy to the random-effect model used in metanalysis, that each of the above terms represents a realization from the same superpopulation. Thus, the assumption is not that the various studies are "identical," but rather that they are "similar.") The fundamental insight is that B_u, R, and S are more likely to be similar across datasets than are the partial regression coefficients.

In order to obtain appropriately calibrated values of $\hat{Y}$, an estimate of the intercept of the above multivariable regression model is also required. This can be obtained by forcing the predicted regression function to pass through the point (X_m, Y_m), where X_m is the vector of mean values of the predictors, and Y_m is the mean response.

Assessment

The fundamental assumption of the univariable synthesis method is that the various first-order summary measures B_u, R, and S are comparable across datasets. More precisely, this fundamental assumption holds that the values of B_u, R, and S, obtained from various candidate datasets, are similar to those values of B_u^*, R^*, and S^* that would have been obtained from the gold-standard dataset, if these data were available. If the above inputs are comparable, then applying (8) for B^{*1Q} to the set of first-order summary measures from the candidate datasets is conceptually equivalent to calculating B^{*1Q} from the gold-standard dataset, and thus to recreating the best possible estimate of the desired gold-standard regression model.

The validity of this basic assumption, and thus of the methodology as a whole, can potentially be assessed in two ways. First, we could ask the *empirical* question, namely, *to what degree do estimates of R, B_u, and S tend to be similar across multiple datasets?* (The question of whether B_u is similar across datasets is a standard problem in metanalysis—the question of whether R and S are similar has been less exhaustively studied.) Second, we

could ask the *mathematical* question, namely, *what is the impact, on the partial regression coefficients and predicted values for individual subjects, of discrepancies between the gold-standard estimates of R^*, B_u^*, and S^* and estimates of R, B_u, and S obtained from the candidate datasets?* In other words, we could perform a mathematical sensitivity analysis to determine the degree to which the above discrepancies in the inputs are likely to affect the outputs.

Both assessment approaches suffer from a fundamental difficulty; namely, that the number of potential regression models to which the proposed technique could be applied is infinite (eg, regression models can differ in the number of predictor variables as well as the values of B_u, R, S, and B). Therefore, (a) demonstrating that the method works well in one circumstance does not necessarily demonstrate that it will work well in others; and (b) the number of possible circumstances is so large that it is difficult to develop a set of scenarios that would be sufficiently representative. We deal with this difficulty by setting up a single scenario (described in detail later) that is both simple and typical. Given this scenario, we then perform a mathematical sensitivity analysis across a wide range of parameter values and observe the effects of these changes on (a) the estimated regression coefficients and (b) set of the predictions generated by the model. Though not intended to be a definitive analysis, this approach does allow us to assess the robustness of the methodology in its most basic form; and also to illustrate how the users of this methodology can set up a sensitivity analysis that is tailored to the characteristics of their own data.

Sensitivity analysis methods

The dataset for the sensitivity analysis has 84 subjects and 3 variables: an outcome Y, a commonly accepted predictor X_1, and a new predictor X_2. (The raw data happened to be taken from a study in exercise physiology, but the source is not as important as the fact that X_1 and X_2 operate in exactly the same fashion as risk factors in epidemiologic investigations.) Table 1 provides a list of the data.

The gold-standard multivariable regression, having $R^2 = 0.67$, is

$$\hat{Y}^* = 1743.94 - 92.65X_1^* + 39.44X_2^*. \quad (9)$$

The standard deviations of b_1^* and b_2^* above are 7.93 and 12.07, respectively. The univariable regressions are

$$\hat{Y}^* = 1751 - 76.53X_1^*, \quad (10)$$

where $R^2 = 0.62$ and

$$\hat{Y}^* = 576.19 - 48.34X_2^*, \quad (11)$$

where $R^2 = 0.11$. The standard deviations of these univariable regression coefficients b_{u1}^* and b_{u2}^* are 6.56 and 15.38, respectively. All of the regression coefficients are statistically significant. The correlation between the predictors is 0.62, and the standard deviations of the predictors are 0.94 and 0.62, respectively. In this dataset (a) the commonly accepted risk factor is a relatively good predictor of the outcome; (b) once the commonly accepted risk factor is included in the model, the new predictor has an

Table 1. Raw data used in simulation examples.

y	x_1	x_2	y	x_1	x_2
223.1	19.8	8.3	149.0	21.1	8.6
105.4	21.0	8.5	171.0	20.3	8.8
161.9	21.4	8.8	111.0	21.4	8.9
161.3	21.3	9.0	99.0	21.8	9.2
94.1	21.0	8.2	267.0	19.0	8.4
280.5	19.7	8.3	98.0	21.0	8.6
183.6	19.7	8.0	184.1	19.0	8.4
204.4	21.0	8.7	416.1	19.0	8.4
140.2	20.2	8.1	112.3	20.5	8.6
73.0	21.4	9.1	583.6	19.0	8.5
194.0	20.4	8.4	53.4	21.8	8.9
118.0	21.1	8.6	180.4	19.8	8.1
68.3	22.0	9.6	128.0	21.6	8.8
131.0	21.4	9.1	82.4	21.7	9.1
127.0	21.0	8.5	230.8	20.4	8.7
72.2	21.6	8.9	135.2	21.6	8.9
93.0	21.8	9.5	90.8	22.0	9.6
94.9	21.0	8.9	181.0	20.5	8.8
108.3	22.0	9.2	99.0	21.7	8.7
118.9	20.3	6.7	321.6	19.0	8.6
83.8	20.9	6.6	134.7	21.5	8.7
66.6	22.0	9.9	342.0	19.9	8.4
117.7	21.1	8.6	115.0	20.9	8.6
209.6	19.0	6.3	185.0	20.9	8.7
137.0	20.8	8.5	164.0	20.0	8.9
66.0	21.2	8.8	89.6	22.0	9.8
174.8	21.1	8.4	225.7	20.5	8.5
427.8	19.0	9.0	179.1	20.7	8.5
179.6	21.4	8.9	54.9	22.0	9.3
237.3	19.5	8.0	96.3	21.5	8.3
209.9	19.8	8.4	71.0	21.2	8.9
319.0	19.3	8.1	62.5	22.0	10.0
89.7	21.6	8.6	191.8	19.5	8.1
122.0	22.0	9.4	65.0	21.9	9.2
112.1	22.0	9.2	201.0	21.2	8.8
131.8	21.5	8.7	116.0	21.0	8.7
80.0	22.0	9.6	191.0	20.3	8.3
87.0	21.5	8.6	136.7	21.1	8.8
247.0	19.0	8.3	137.4	21.1	8.8
70.0	21.1	9.2	67.0	22.0	10.0
63.5	22.0	9.3	207.0	19.4	8.4
224.7	20.4	8.7	122.0	21.3	9.3

incremental benefit which is of moderate magnitude; (c) the commonly accepted and new risk factors are positively correlated; and (d) when comparing the multivariable and univariable models, some of the parameter values differ (indeed, the regression coefficient for X_2^* changes sign). These characteristics are present in many epidemiological datasets.

To implement the sensitivity analysis, we modified three of the inputs: (a) the values of R^* were varied by adding from −0.10 to +0.10, in increments of 0.01, to the baseline value of 0.62; (b) the values of B_u^* were varied by adding from −15 to +15, in increments of 1.5, to the baseline values of −76.54 and −48.34; and (c) the values of S^* were varied by adding from −0.15 to +0.15, in increments of .015, to the baseline values of 0.94 and 0.62. The differences between these inputs and the true values from the gold-standard dataset play the role of the variability likely to be observed by using the candidate datasets rather than the gold-standard dataset. (The above perturbations of the inputs were derived on intuitive grounds in order to represent from small to moderately large differences between the above datasets—for example, the extreme values for B_u^* are in the range of 1–2 standard deviations from the values in the gold-standard dataset. In practice, the user might base the choice of perturbations on more substantive considerations pertinent to the scientific issues at hand.)

For each set of simulation inputs, we reestimated the multivariable regression model using (8), thus obtaining the following: (a) new multivariable regression coefficients and (b) new predicted values. To determine how close the new multivariable regression coefficients were to the gold-standard values, we calculated a standardized distance (D) [30]:

$$D = \left\{ \frac{\{[(b_1 - b_1^*)/s(b_1)]^2 + [(b_2 - b_2^*)/s(b_2)]^2\}}{2} \right\}^{1/2}. \tag{12}$$

For example, for the simulation with B_u unchanged, S unchanged, and R increased from 0.62 to 0.66, the estimated partial regression coefficients become −98.87 and 51.36. The standardized distance is

$$\begin{aligned} D &= \left\{ \frac{1}{2} \left(\left[\frac{6.22}{6.56} \right]^2 + \left[\frac{11.92}{15.38} \right]^2 \right) \right\}^{1/2} \\ &= (0.75)^{1/2} = 0.87 \end{aligned} \tag{13}$$

implying that the average change in the partial regression coefficients is a bit less than one standard deviation. To determine how consistent the predicted values were, we took the correlation between $\hat{Y}$ and $\hat{Y}^*$, where $\hat{Y}$ and $\hat{Y}^*$ are the vectors (ie, across all subjects) of predicted outcomes for the two models in question. For the above example, the correlation was 0.997.

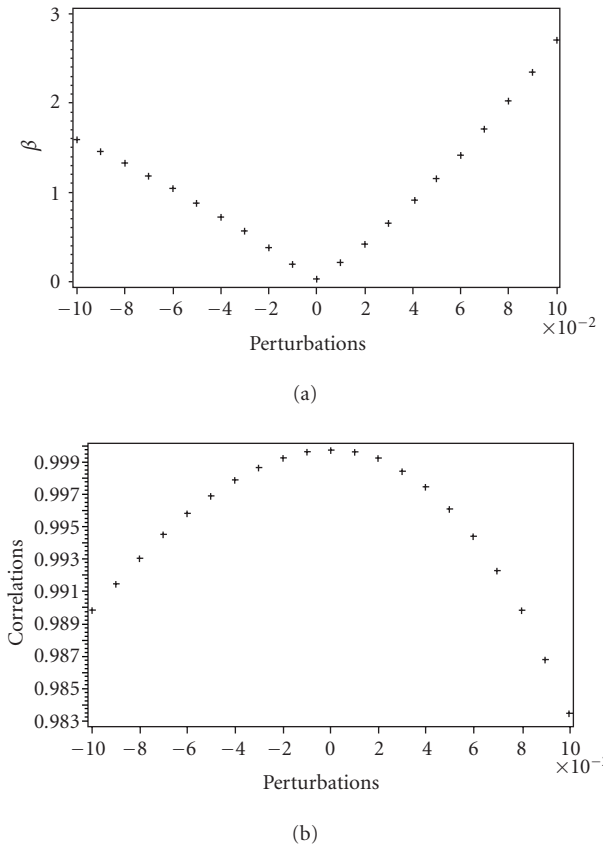


Figure 1. (a) Univariable synthesis method—effect of perturbing R on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing R on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

RESULTS

Figures 1–5 summarize the results. In particular, each set of figures describes the impact, on either the standardized difference between B and B^* (Figures 1a, 2a, and 3a) or the correlation between $\hat{Y}$ and $\hat{Y}^*$ (Figures 1b, 2b, and 3b), of perturbing one of the inputs, while keeping all other inputs at the true values from the gold-standard dataset. Figure 1 shows the effects of perturbing R. Figure 2 shows the effects of perturbing b_{u2}. Similar results were found for perturbing b_{u1}. Figure 3 shows the effects of perturbing s_1. Similar results were found for perturbing s_2.

Figures 4 and 5 show the effects of perturbing both b_{u1} and b_{u2} on the estimated values of $Y(\hat{Y})$ and on the model residuals compared to the unperturbed model. Similar results were found for perturbing both s_1 and s_2. The residuals from the perturbed models had similar distributions

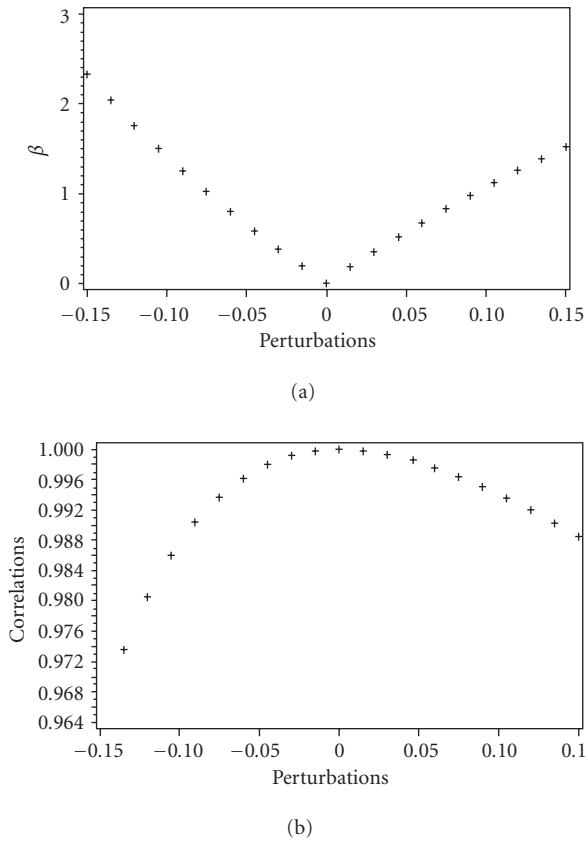


FIGURE 2. (a) Univariable synthesis method—effect of perturbing b_{u2} on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing b_{u2} on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

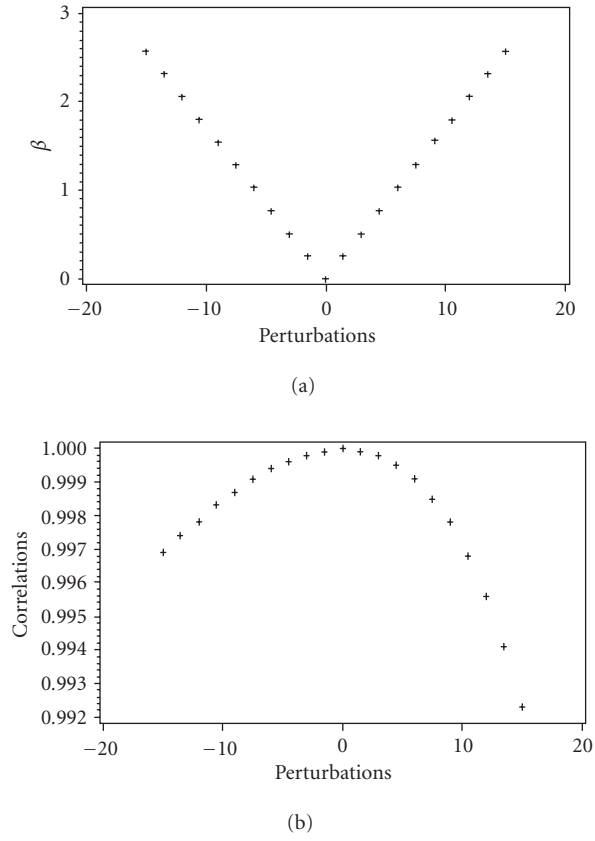


FIGURE 3. (a) Univariable synthesis method—effect of perturbing s_1 on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing s_1 on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

compared to those of the unperturbed model (data not shown). Also, plots of the residuals from the perturbed and unperturbed models against X_1, X_2, and $\hat{Y}^*$ were very similar (data not shown).

Even modest perturbations of the inputs affect the estimated values of the partial regression coefficients; for example, varying R by 0.05 units is associated with an approximately 1-unit difference between B and B^*. Perturbing the inputs has much less impact on the correlation between the predicted values. For example, applying the above perturbation to R resulted in a correlation between $\hat{Y}$ and $\hat{Y}^*$ exceeding 0.99. Similar results were observed when perturbing all the inputs simultaneously (data not shown).

In summary, the univariable synthesis approach appears to be robust to changes in its inputs, so long as what the user is ultimately interested in is the predicted values resulting from the multivariate regression. The methodology is relatively less robust when estimating the values of the partial regression coefficients.

DISCUSSION

Creating multivariable regression models containing partial regression coefficients is central to the practice of epidemiology. It is quite common for the risk factors (predictors) of interest to be distributed across multiple datasets. Because the value of partial regression coefficients depends upon the choice of the other variables that are included in the model, simply combining partial regression coefficients across datasets may be dangerous. Indeed, combining partial regression coefficients across datasets is the most dangerous in the situation of most practical interest, that is, when the correlations among the risk factors in question are moderate to strong. One strength of the univariable synthesis method is that the

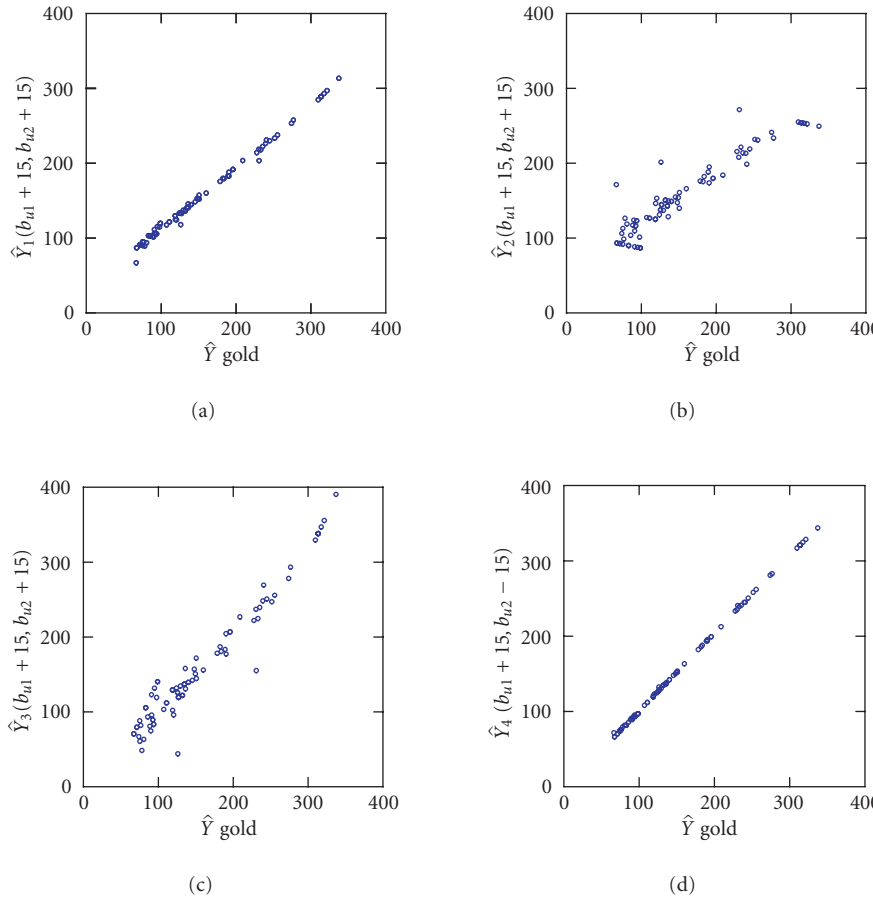


FIGURE 4. Univariable synthesis method—effect of perturbing both b_{u1} and b_{u2} on correlations between the predicted values for the perturbed and unperturbed models. The x-axes represent the estimated Y of the unperturbed model ($\hat{Y}$ gold); the y-axes represent the estimated Y of the perturbed models. The y-axes labels indicate the perturbation. For example, for Y_1, both b_{u1} and b_{u2} were perturbed by adding 15. The estimating equation was then computed and $\hat{Y}_1$ was calculated.

correlations among the predictors are explicitly considered in the quantitative estimation of the partial regression coefficients.

We know of no ideal solution to this problem, but have proposed the univariable synthesis method as a possible way forward. The most critical assumption underlying this method is that first- and second-order information such as univariable regression coefficients, standard deviations, and correlations are comparable across datasets. Admittedly, the assumption of comparability is strong, but it is not essentially different from what must be assumed in order to make qualitative conclusions about epidemiological phenomena based on information from multiple sources, or what must be assumed when information about individual risk factors is quantitatively combined across studies using metanalysis. In any event, it might be argued that (a) these assumptions are being made explicitly rather than implicitly; (b) sensitivity analyses can be performed in order to assess the impact of these assumptions; and (c) the alternatives—namely, ignoring the issue entirely or limiting the number of risk factors to be modeled—have significant difficulties of their own.

The univariable synthesis method has a number of limitations. As discussed above, it assumes that first- and second-order information can be combined across datasets. Fortunately, the technique appears to be reasonably robust to modest departures from these assumptions—particularly when the focus of inference is on the predictions generated by the model rather than the parameter estimates themselves. Other limitations include the inability to deal with interactions and the difficulty of generating estimates of precision (eg, standard errors of regression coefficients).

A limitation of our assessment is the less-than-comprehensive nature of the sensitivity analyses. In essence, by selecting a single dataset to use as an archetype, we implicitly assume that the goal of the sensitivity analyses is demonstration of the plausibility of the concept rather than definitive proof. This is a generic problem in

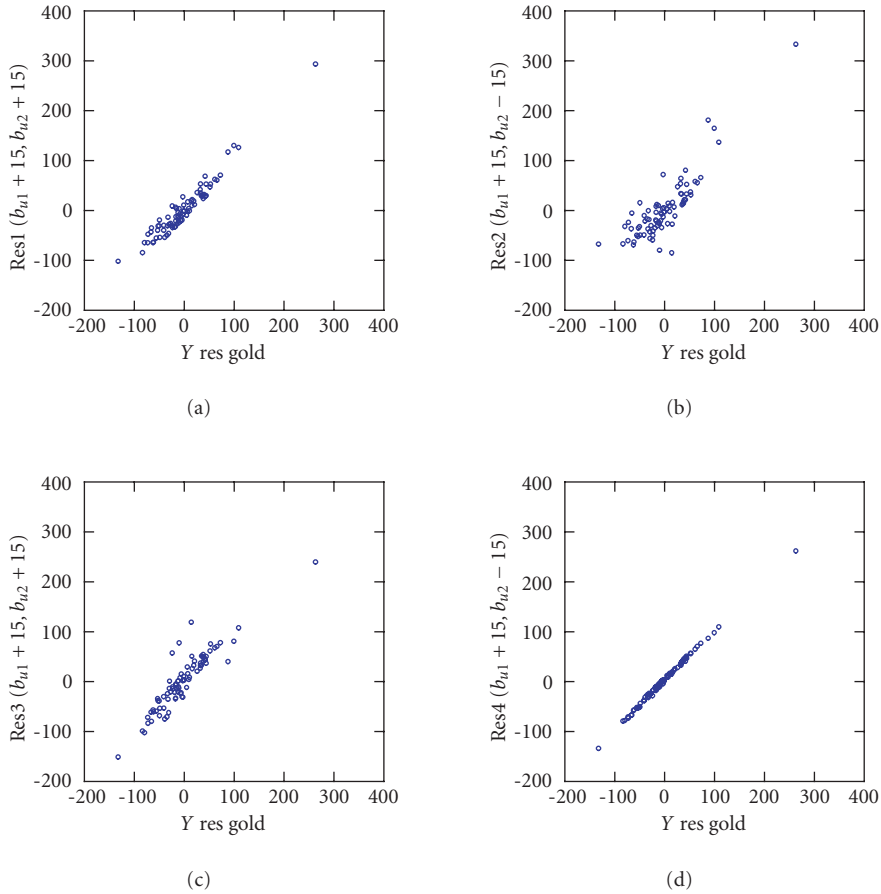


FIGURE 5. Univariable synthesis method—effect of perturbing both b_{u1} and b_{u2} on residuals of the perturbed and unperturbed models. The x-axes represent the residuals of the unperturbed model (Y res gold); the y-axes represent the residuals of the perturbed models. The y-axes labels indicate the perturbation. For example, for RES1, both b_{u1} and b_{u2} were perturbed by adding 15. The estimating equation was then computed and Y_1 residual was calculated.

the use of simulation methodology to analyze the properties of statistical methods having application across a wide range of conditions.

An implication of the above is that before using the univariable synthesis method in practice, the user should always perform a sensitivity analysis relevant to his or her application. The observed data should be assumed to represent the gold-standard, and the implications of permuting the inputs to the synthesis analysis techniques can be assessed as illustrated here. (Thus, a further assumption is being made—namely, that the local behavior of the system near the values of the gold-standard estimates can be adequately modeled by the local behavior of the system near the sampled values from the candidate datasets.)

A final limitation applies to those applications where the regression coefficients are of more interest than the predicted values. The univariable synthesis method is more robust with respect to its predicted values than to the values of its regression coefficients. In large part, this may simply be a reflection of the general instability of partial regression coefficients.

Under what circumstances might the univariable synthesis method be applied? Perhaps the most natural application would be to generate lists of patients at high-risk. For example, a predicted length of stay for post-stroke rehabilitation could be generated, the 10% of patients with the highest predicted lengths of stay could be identified, then be targeted for an intervention intended to reduce this length of stay. Such an application focuses much more on predicted values than regression coefficients, and thus makes use of the component of this methodology with the greatest apparent robustness. In this case, the interpretation of the simulation results indicates that the correlation between the gold-standard and the candidate datasets becomes critical. For example, (assuming a normal distribution of predicted values) if this correlation is 0.95, 0.97, and 0.99, then, of those patients with the highest 10% of predicted values generated by the univariable synthesis methodology, approximately 79%, 83%, and 91% of patients will be in the top 10% generated from the gold-standard database. (If the distribution of predicted values has heavier tails than the normal distribution, then

these percentages will be even higher.) Thus, the magnitude of correlations observed in our simulations implies that those patients identified by the univariable synthesis method as having extreme predicted values of the outcome are likely to be actually extreme.

One principle that is implicit in the above discussion is that, because the univariable synthesis method is assumption intensive, in any given circumstance its application will involve a trade-off between its approximate nature (a negative) and the improvement in prediction obtained by being able to include additional risk factors (a positive). Thus, this trade-off would be most likely to favor the adoption of the new method in situations where (a) substantive considerations suggest that the various candidate datasets are comparable (ie, thus reducing the negative impact of the assumptions); (b) the new predictors explain a substantively important amount of the variation outcome, above and beyond the traditional predictors (ie, thus, increasing the positive impact of being able to include new predictors); and (c) the primary focus is on the predicted values themselves rather than the models' partial regression coefficients (ie, because the robustness of the method is greatest for their predicted values). Encouragingly, these conditions describe a significant area of epidemiological practice, especially if the set of potential outcomes is expanded to include dichotomous outcomes (such as the incidence of disease) and time until survival. Extensions of the univariable synthesis and related methods to other types of outcomes will be presented elsewhere. Ongoing challenges for the developers involve both extending these methods and determining the set of applications for which these new tools are best suited.

APPENDICES

A SAS implementation of univariable simulation method

The input file, named inputs, contains x_0, x_1, x_2, and y;

```
proc iml;
  use inputs (keep=y);
  read all  into y;

  use inputs (keep=x_0);
  read all into x0;

  use inputs (keep=x_1);
  read all into x1;

  use inputs (keep=x_2);
  read all into x2;

  x1x2=x1||x2;
  x0x1=x0||x1;
  x0x2=x0||x2;
  x=x0||x1||x2.
```

This portion of the code generates the standard regression results;

```
  xpxi=inv(t(x)*
  x);
  beta=xpxi*(t(x)*y);
  yhat=x*beta;
  resid=y-yhat;
  sse=ssq(resid);
  n=nrow(x);
  dfe=nrow(x)-ncol(x);
  mse=sse/dfe;
  cssy=ssq(y-sum(y)/n);
  rsquare=(cssy-sse)/cssy;
  r=corr(x1x2);
  stdb=sqrt(vecdiag(xpxi)*mse);

  beta1=inv(t(x0x1)*x0x1)*(t(x0x1)*y);
  beta2=inv(t(x0x2)*x0x2)*(t(x0x2)*y);
```

This portion of the code implements the univariable synthesis approach;

```
  s1=sqrt((ssq(x1-sum(x1)/n))/(n-1));
  s2=sqrt((ssq(x2-sum(x2)/n))/(n-1));

  bu=beta1[2,1] // beta2[2,1];
  s=s1 // s2;
  invr=inv(r);
  bus=bu#s;
  invrbus=invr*bus.
```

b_syn is the estimated regression coefficient, yhat_syn is the predicted outcome, where yhat_syn can be further modified to lie on the line with slope b_syn and passing through the point consisting of the means of all variables;

```
  b_syn=(inv(r)*(bu#s))/s;
  yhat_syn=x1x2*b_syn;


print b_syn yhat_syn;

quit;
 run.
```

B Illustration of the calculations for the univariable synthesis method

From the dataset in Table 1,

$$R = \begin{bmatrix} 1.0000000 & 0.6223841 \\ 0.6223841 & 1.0000000 \end{bmatrix},$$

$$B_u = \begin{bmatrix} -76.52528 \\ -48.34035 \end{bmatrix},$$

$$S = \begin{bmatrix} 0.9403281 \\ 0.6177001 \end{bmatrix}.$$

The steps in the calculation are as follows:

$$R^{-1} = \begin{bmatrix} 1.6322853 & -1.015908 \\ -1.015908 & 1.6322853 \end{bmatrix},$$

$$B_u \cdot S = \begin{bmatrix} -71.95886 \\ -29.85984 \end{bmatrix},$$

$$R^{-1}(B_u \cdot S) = \begin{bmatrix} -87.12253 \\ 24.363844 \end{bmatrix},$$

$$(R^{-1}(B_u \cdot S))/S = \begin{bmatrix} -92.65121 \\ 39.442839 \end{bmatrix}.$$

ACKNOWLEDGMENT

This research was funded by BioSignia Inc, which placed no limitations on publication.

REFERENCES

[1] Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987;9:1–30.

[2] Egger M, Schneider M, Smith GD. Spurious precision? Meta-analysis of observational studies. *BMJ.* 1998;316(7125):140–144.

[3] Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol.* 1999;28(1):1–9.

[4] Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol.* 1994;140(3):290–296.

[5] Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). *Ann. Rev. Public Health.* 1996;17:1–23.

[6] Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ.* 2001;323(7304):101–105.

[7] Ernst E, Resch KL. Fibrinogen as a cardiovascular risk factor: a meta-analysis and review of the literature. *Ann Intern Med.* 1993;118(12):956–963.

[8] Etminan M, Gill S, Samii A. Effect of non-steroidal anti-inflammatory drugs on risk of Alzheimer's disease: systematic review and meta-analysis of observational studies. *BMJ.* 2003;327(7407):128–132.

[9] Vincent JL, Dubois MJ, Navickis RJ, Wilkes MM. Hypoalbuminemia in acute illness: is there a rationale for intervention? A meta-analysis of cohort studies and controlled trials. *Ann Surg.* 2003;237(3):319–334.

[10] Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. *JAMA.* 1998;279(18):1477–1482.

[11] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med.* 2002;21(4):589–624.

[12] Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Stat Med.* 2003;22(14):2309–2333.

[13] Arends LR, Vokó Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. *Stat Med.* 2003;22(8):1335–1353.

[14] Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA.* 1997;278(16):1349–1356.

[15] Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. *JAMA.* 2004;291(2):210–215.

[16] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst.* 1989;81(24):1879–1886.

[17] Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst.* 1999;91(18):1541–1548.

[18] Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* 1998;90(18):1371–1388.

[19] Anderson S, Ahnn S, Duff K. NSABP Breast Cancer Prevention Trial Risk Assessment Program (Version 2). NSABP Biostatistical. *Center Technical Report.* 1992.

[20] Anderson KM, Wilson PW, Odell PM. An updated coronary risk profile. A statement for health professionals. *Circulation.* 1991;83(1):356–362.

[21] Laurier D, Nguyen PC, Cazelles B, Segond P., Estimation of CHD risk in a French working population using a modified Framingham model. The PCV-METRA Group. *J Clin Epidemiol.* 1994;47(12):1353–1364.

[22] Menotti A, Lanti M, Puddu PE, Kromhout D. Coronary heart disease incidence in northern and southern European populations: a reanalysis of the seven countries study for a European coronary risk chart. *Heart.* 2000;84(3):238–244.

[23] D'Agostino R, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA.* 2001;286(2):180–187.

[24] Marrugat J, D'Agostino R, Sullivan L. Adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. *J Epidemiol Community Health.* 2003;57(8):634–638.

[25] Matchar DB, Samsa GP, Matthews JR, et al. The Stroke Prevention Policy Model: linking evidence and clinical decisions. *Ann Intern Med.* 1997;127:704–711.
[26] Samsa GP, Reutter RA, Parmigiani G, et al. Performing cost-effectiveness analysis by integrating randomized trial data with a comprehensive decision model: application to treatment of acute ischemic stroke. *J Clin Epidemiol.* 1999;52(3):259–271.
[27] Pigott TD. Missing predictors in models of effect size. *Eval Health Prof.* 2001;24(3):277–307.
[28] Zhao LP, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equation. *Biometrics.* 1996;52(4):1165–1182.
[29] Steyerberg EW, Eijkemans MJ, van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Stat Med.* 2000;19(2):141–160.
[30] Draper N, Smith H. *Applied Regression Analysis.* 2nd ed. New York, NY: John Wiley & Sons. 1981.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 124–131 • DOI: 10.1155/JBB.2005.124

RESEARCH ARTICLE

Contextual Multiple Sequence Alignment

Anna Gambin and Rafał Otto

Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland

Received 30 November 2003; revised 27 February 2004; accepted 12 March 2004

In a recently proposed contextual alignment model, efficient algorithms exist for global and local pairwise alignment of protein sequences. Preliminary results obtained for biological data are very promising. Our main motivation was to adopt the idea of context dependency to the multiple alignment setting. To this aim the relaxation of the model was developed (we call this new model *averaged contextual alignment*) and a new family of amino acids substitution matrices are constructed. In this paper we present a contextual multiple alignment algorithm and report the outcomes of experiments performed for the BAliBASE test set. The contextual approach turned out to give much better results for the set of sequences containing orphan genes.

INTRODUCTION

The multiple alignment of biological sequences has become an essential tool in molecular biology. It is used to find conserved regions and motifs in protein families, to detect the homology between new sequences and groups of sequences having an already known function and in a preliminary phase of protein structure prediction. Multiple alignment is also extensively used in molecular evolutionary analysis.

The various genome projects have provided the biologist with a great number of new protein sequences, and the rate of appearance of these data is steadily increasing. The development of an accurate and reliable multiple alignment program which is capable of handling many (often very divergent) sequences simultaneously is still of major importance.

The complexity of the problem does not allow to find the exact solution in a reasonable computational time [1]. Traditionally, the most popular heuristic approach has been the progressive alignment method [2].

In this paper we propose to explore new model for sequence alignment, in which the score for the substitution also depends on its neighborhood in the sequence. Such *contextual alignment model* has been proposed recently in [3] for the pairwise alignment problem. Preliminary results obtained for biological data by Gambin and Slonimski in [4] are very promising. To apply the contextual approach in the multiple alignment setting we have decided to relax slightly the model from [3]. However, we still need the family of contextual amino acid substitution matrices, for which a novel construction procedure is described. We present preliminary experimental results that illustrate the advantage of using a contextual approach in progressive alignment algorithm. It turned to be particularly useful in aligning the family of sequences containing several *orphans* (these are distantly related sequences, sometimes sharing the common fold).

It should be clear that the existence of orphan genes is unavoidable. Despite the accumulation of genetic information, newly sequenced genomes continue to reveal a high proportion (even to 50%) of uncharacterized genes. Among them there is a significant number of strictly orphan genes without any resemblance to previously determined protein sequences. Moreover, most genes found in databases have only been predicted by computer methods and have never been experimentally validated. Hence, for the alignment method it is important to tolerate orphans (some existing programs exclude the divergent orphans as unrelated or unalignable sequences) and to keep the stability of the family alignment when orphans are introduced into the sequence set.

The paper is organized as follows. We start with the description of an averaged contextual model. Then we present a construction method for contextual substitution tables. The next section proposes the progressive multiple alignment algorithm that takes context into account. The results of experimental analysis are presented in "results," which is followed by conclusions and discussion of further works.

Correspondence and reprint requests to Rafał Otto, Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Email: Poland; rotto@mimuw.edu.pl

AVERAGED CONTEXTUAL MODEL

The contextual alignment model considered in [3] cannot be directly applied to the problem of multiple alignment. In this model the score of an alignment depends on the order of operations (substitutions and indels) performed, as a substitution at one position can change the context for neighboring sites. The optimal alignment for the pair of sequences was defined as the alignment having the maximal score, when we maximize over all possible chronologies of evolutionary changes. More detailed study of the structure of optimal alignments and the description of efficient algorithms constructing them are included in [3].

To deal with several sequences simultaneously and also to keep the context dependency, we propose a relaxed contextual model. In this model we also penalize substitution considering two surrounding letters but we do not take care of the relative order of operations. In our algorithm the context independent and affine gap penalties are assumed.

Consider the following example of a short fragment of pairwise alignment:

$$\begin{aligned} &\ldots \text{HCA} \ldots \\ &\ldots \text{ADG} \ldots \end{aligned}$$

In the contextual model the score for substitution $C \mapsto D$ would depend on the order of operations. For instance, if the substitution $H \mapsto A$ has been performed after the substitution $C \mapsto D$ and the substitution $A \mapsto G$ has been performed before $C \mapsto D$, then the substitution $C \mapsto D$ would have the left context H and the right context G. In our simplified model we consider all 4 possible contexts for the middle substitution and take an average of 4 contextual scores. Notice that standard noncontextual, for example, Blosum matrix entry can be viewed as an average over all 400 possible pairs of contexts.

As a second example, consider the substitution surrounded by a deletion on the left and an insertion on the right:

$$\begin{aligned} &\ldots \text{HAHCC} - - - \text{A} \ldots \\ &\ldots \text{A} - - - \text{DDGAG} \ldots \end{aligned}$$

Now we have 9 possible contexts for the substitution $C \mapsto D$. On the left there are two different operations: the substitution $H \mapsto A$ and the deletion of AHC. If none of them has happened before the substitution $C \mapsto D$, then the left context is C. If AHC has been deleted before, then the left context is A or H depending on the relative order of these two substitutions. Analogous cases can be considered for the right context of the substitution $C \mapsto D$. As before we count the score as the average over 9 contextual scores.

CONTEXTUAL SUBSTITUTION TABLES

Methods

The contextual alignment algorithm, as an important part of its input data, takes a contextual scoring table, which provides the score for every possible substitution in every possible context.

The family of matrices proposed in [5] suffers from the fundamental difficulty that the amount of data necessary to construct a complete contextual substitution table exceeds the data presently available by an order of magnitude. To cope with this problem we present here a new approach to the construction of contextual substitution tables. The algorithm is in fact a contextual extension of the one that has been used to create Blosum tables [7].

As the input data we take the database of blocks (ungapped fragments of multiple alignments) and start by computing the observed frequency of substitutions. The extension from the existing method is the fact that we distinguish substitutions having different contexts. Let $f_{i,j}^{k,l}$ denote the number of observed substitutions $i \mapsto j$ in the context of k and l.

Each of the considered substitutions can have 4 different contexts so instead of increasing by 1 the entry $f_{i,j}$ we increase entry $f_{i,j}^{k,l}$ by 1/4 for all four possible pairs of (k, l).

Having computed frequency table we define the observed frequency for each substitution $i \mapsto j$ in the context (k, l) as

$$q_{i,j}^{k,l} = \frac{f_{i,j}^{k,l}}{\sum_{i,j} \sum_{k,l} f_{i,j}^{k,l}}. \tag{1}$$

Now, we can compute the observed frequency of the residue i in the context (k, l) as

$$p_i^{k,l} = q_{i,i}^{k,l} + \frac{1}{2} \sum_{i \neq j} q_{i,j}^{k,l}, \tag{2}$$

and the observed frequency of the context (k, l) as $u^{k,l} = \sum_i p_i^{k,l}$.

The expected frequency of the substitution $i \mapsto j$ in the context (k, l) is given by

$$e_{i,j}^{k,l} = \begin{cases} \dfrac{p_i^{k,l} p_j^{k,l}}{u^{k,l}} & \text{for } i = j, \\ \dfrac{2 p_i^{k,l} p_j^{k,l}}{u^{k,l}} & \text{for } i \neq j. \end{cases} \tag{3}$$

Finally the score for $i \mapsto j$ in the context (k, l) is

$$s_{i,j}^{k,l} = \log_2 \left(\frac{q_{i,j}^{k,l}}{e_{i,j}^{k,l}} \right). \tag{4}$$

To avoid the influence of highly similar sequences we adopt the idea of clustering inside blocks as it was done in the Blosum table.

Table 1. Characteristics of substitution scores.

Clustering %	NONCTX			CTX		
	Avg	StdDev	Entropy	Avg	StdDev	Entropy
100%	−0.5984	1.4561	1.0642	−0.4715	1.5498	1.0558
90%	−0.3363	1.2165	0.6515	−0.3124	1.2537	0.6620
80%	−0.2472	1.1086	0.5128	−0.2310	1.1316	0.5248
70%	−0.1658	0.9878	0.3839	−0.1590	0.9999	0.3970
60%	−0.0928	0.8449	0.2590	−0.0931	0.8523	0.2716
50%	−0.0429	0.6858	0.1519	−0.0500	0.7040	0.1622
40%	−0.0110	0.5607	0.0883	−0.0278	0.6013	0.1002

Table 2. The robustness of noncontextual tables. The range, median, and standard deviation for the number of examples drawn on per substitution score.

Table	No of pairs used	Min	Max	Med	StdDev
NONCTX100	910427386	201204	44246771	2089431	6447364
NONCTX90	397939179	85159	17134863	1154422	2226931
NONCTX80	228719630	52834	8703468	719781	1159503
NONCTX70	125188080	32428	4022674	429833	563942
NONCTX60	58669007	17718	1468427	218982	228104
NONCTX50	21889157	8121	424847	94091	70724
NONCTX40	7104342	3034	110252	32151	21160

Table 3. The robustness of contextual tables. The range, median, and standard deviation for the number of examples drawn on per substitution score.

Table	No of pairs used	Min	Max	Med	StdDev
CTX100	910427386	80	3033544	72832	105276
CTX90	397939179	52	1112640	38208	33440
CTX80	228719630	40	504100	21952	17628
CTX70	125188080	24	211316	14016	8620
CTX60	58669007	8	98096	1644	3560
CTX50	21889157	4	18736	700	1116
CTX40	7104342	1	7016	228	352

Results

As an input we have taken the BLOCKS+ database available at http://blocks.fhcrc.org (see [6]), which consists of 11 858 blocks representing 2608 groups. We have derived two kinds of tables: noncontextual (using the method in [7]) and contextual using the method described above. We have created tables with 7 different clustering percentages: 100%, 90%, 80%, 70%, 60%, 50%, and 40%. In Table 1 several characteristics of computed tables are summarized. The interesting observation is that the contextual tables have higher average score and higher entropy. Entropy also increases with clustering percentage as a normal consequence of reducing multiple contributions to amino acid pair frequencies from the most closely related sequences in the block. For the discussion of the notion of entropy in the context of substitution tables see [8].

The size of contextual tables (84 000 entries instead of 210 in case of noncontextual tables) implies a small amount of data that supports each table entry. If these statistics were too low, this could have direct impact on the quality of the score value. Tables 2 and 3 give a good view of these issues. Therefore we should discuss the robustness of proposed methods. The substitution table which was finally used in our experiments (CTX70) has an acceptable number of pairs impacting the average score; moreover there was no hole (blank entry) in this table.

For interested readers the matrices parameterized by different clustering constants can be found at http://www.mimuw.edu.pl/~aniag/TABLES.

MULTIPLE ALIGNMENT ALGORITHM

The averaged contextual model is proposed to enable computing multiple alignment in the contextual manner.

Ignoring the relative order of operations in the alignment simplifies the task of computing multiple alignment; however it is still not easy to keep the complexity on the reasonable level. Multiple alignment dynamic programming algorithm is extremely time consuming even in the case of the noncontextual model. A lot of heuristic approaches, which have been already developed, try to reach the optimal solution with the highest possible probability. The progressive alignment [1, 9] is one of the most popular alignment approaches.

Our contextual multiple alignment algorithm can be viewed as a contextual extension of popular ClustalW algorithm [9] or Feng-Doolittle algorithm [2], which belong to the family of progressive alignment algorithms. The main idea is to align pairs of sequences progressively and to deduce the multiple alignment from the set of pairwise alignments.

To this aim we have developed efficient averaged contextual pairwise alignment algorithms. These are appropriately modified standard dynamic programming procedures. We omit the details here; for interested readers all algorithms implemented in C++ can be found at http://www.cern.ch/rotto/Biology/Sources/ACM.

The important remark here is that we do not (not yet) intend to concur with the existing algorithms. Our goal is to demonstrate the usefulness of the contextual approach. We want to design algorithms that can be applied to the contextual model as well as to the noncontextual one. Then, we are able to compute alignments in both models and finally compare the results. In fact the ClustalW algorithm is equipped with the huge number of additional nontrivial heuristics (such as sequences weighting, substitution matrices varied at different alignment stages, residue-specific gap penalties, etc) which are not applied in our algorithm.

An overview of our algorithm is as follows:

(1) Calculate a distance matrix from pairwise scores for a given group of sequences.

(2) Construct a *guide tree* from the distance matrix using the neighbor-joining clustering algorithm [10].

(3) Progressively align the sequences in order of decreasing similarity. Three kinds of alignments are considered here:

  (i) pairwise alignment of two sequences,

  (ii) alignment of a sequence with an alignment,

  (iii) alignment of two alignments.

Calculating distance matrix

Several methods to derive the pairwise evolutionary distance (sometimes called difference score) from alignment scores are proposed (see, eg, [2]). Being aware of the drawbacks of all these approaches (see Gonnet and Korostensky, *Optimal scoring matrices for estimating distances between aligned sequences* available at http://www.inf.ethz.ch/personal/gonnet/papers/Distance/Distance.html for a detailed discussion) we decided to use the method proposed by Feng and Doolittle [2]. It works for global and local alignments. Assuming that $S(V, W)$ is the local similarity score between the sequences V and W, then their distance is defined via

$$D(V, W) = -\ln\left(\frac{S(V, W) - S_{\text{rand}}}{S_{\text{iden}} - S_{\text{rand}}}\right), \tag{5}$$

where S_{iden} is the average of the two scores for the two sequences compared with themselves and S_{rand} is the expected score of two random sequences with the same amino acids composition as V and W.

Constructing guide tree

The next step in progressive multiple alignment is building the guide tree. Here we use the neighbor-joining method [10] and two alternative methods to find the root of the tree. The first one is by adding an outgroup sequence to the given sequence group [1] and the second is by finding the middle point of the tree.

Aligning

The last step is to progressively align sequences according to the order given by the guide tree. It means that for each internal node of the tree we align sequences already aligned from the left child of the node with sequences already aligned from the right child of the node. In the simplest case this alignment is pairwise, but closer to the root of the tree we have to align two alignments. We decided to solve that problem using the method of Feng and Doolittle [2]. First, in two given alignments we replace gap letter with *neutral letter* X having at the end two groups of sequences over the extended alphabet $\Sigma \bigcup \{X\}$ (where Σ is an alphabet of amino acids). Also, we extend substitution matrix by adding scores for $a \mapsto X$ equal to 0 for all a. Then for each pair V, W of sequences where V is a sequence from the first group and W is a sequence from the second group, we compute pairwise alignment. The alignment with the maximal score is chosen and according to it we align two groups of sequences. Finally, in all sequences we replace *neutral letter* X back with a gap letter. In that way we obtain multiple alignment while reaching the root of the tree.

RESULTS

BAliBASE: multiple alignment test set

BAliBASE (Benchmark Alignments dataBASE) is a database of manually refined multiple sequence alignments available at http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/.

It is specifically designed for the evaluation and comparison of multiple sequence alignment programs. The sequences included in the database are selected from alignments in structural databases (such as FSSP and

HOMSTRAD) or from manually constructed structural alignments taken from the literature. In our experimental analysis we have used the test sequences from 4 (out of 8) parts of the database, the so-called reference sets.

(i) *Reference 1.* It consists of families of equidistant protein sequences of similar length. Sequences are divided into 6 groups depending on theirs lengths and percent residue identity (% ID).

(ii) *Reference 2.* Each set here contains the group of closely related sequences (more than 25% ID) and up to three orphan genes (ie, genes sharing the common fold with the family, but having weak sequence similarity).

(iii) *Reference 3.* It includes groups consisting of several divergent protein families of equidistant sequences. The reference alignments consist of up to 4 families, with less than 25% ID between any two sequences from different families.

(iv) *Reference 6.* This includes the protein families containing repeats of different residue similarity.

Methodology

In the first stage of our experiment the multiple alignments were calculated for all reference sets in two settings: contextual and noncontextual. Then, the results obtained were compared with the reference alignments from the database. For this comparison the following measure (*sum-of-pairs score* [11]) was used. Let A_1 and A_2 be two multiple alignments of N sequences. Denote by M_1 and M_2 the lengths of these multiple alignments. Let $A(i, j)$ stand for the ith residue in the jth sequence of A. Define for two residues a and b $\delta(a, b) = 1$ if and only if $a = b$ and $\delta(a, b) = 0$ if and only if $a \neq b$. Now, for one column from the multiple alignment A we define

$$S(A, i) = \sum_{j=1}^{N} \sum_{k=1, k\neq j}^{N} \delta\big(A(i, j), A(i, k)\big). \tag{6}$$

And, finally

$$\mathrm{SPS}(A_1, A_2) = \frac{\sum_{i=1}^{M_1} S(A_1, i)}{\sum_{i=1}^{M_2} S(A_2, i)}. \tag{7}$$

SPS is the frequency of properly aligned pairs of residues with respect to the reference alignment.

In the second phase of the analysis we examine the robustness of the alignment to the introduction of orphans. To this aim we use the alignments from Reference 2, which contains related families with divergent, orphan sequences. Denote by $\mathcal{G}$ the set of all sequences from the considered group. Let $\phi(\mathcal{G})$ be the subset of $\mathcal{G}$ consisting only of the family of highly related sequences. Firstly, the multiple alignments $A_{\mathcal{G}}$ were calculated for all groups $\mathcal{G}$; then the multiple alignments for reduced groups (without orphans) $A_{\phi(\mathcal{G})}$. Let $\Phi(A_{\mathcal{G}})$ be the operation of cutting out from $A_{\mathcal{G}}$ the rows which correspond to the orphans. Define the following measure:

$$\mathrm{SPS}' = \mathrm{SPS}\big(A_{\phi(\mathcal{G})}, \Phi(A_{\mathcal{G}})\big). \tag{8}$$

It tests the ability of a model to align divergent sequences and also the degree to which the alignment of the family is disrupted by the introduction of the orphans. We have performed the experiments with various substitution tables and gap penalties. The best scores are obtained for NONCTX70 and CTX70 with gap open penalty 5 and gap extension penalty 1.

Results of the first experiment are presented in Table 4. The entries are the SPS measures averaged over the groups of sequences. Clearly, the contextual approach yields much better in case of sequence families from Reference 2 set, which contains families with orphan genes (especially in case of families of short sequences).

Table 5 summarizes the outcomes of the second experiment. The entries here are SPS′ values for investigated groups of sequences. Results of this experiment confirm the observation taken in the previous one.

The advantage of the contextual approach in aligning families containing orphan genes shown above is quite clear. Here we present some statistics taken on the whole set of experimental data to show that our method is performing a little better than other existing methods also in general case. Figure 1a proves that the results of the contextual model fit those given by the noncontextual one, but Figure 1b shows that contextual scores are more uniform. The fact that the contextual approach improves alignment more significantly in case of small values of noncontextual score is presented in Figure 1c.

Figure 1, Tables 4 and 5 then show the contextual model performs slightly better than the noncontextual one. However, there are some examples when the contextual approach yields much better results. Among them we have the families listed in Table 6.

The challenging task here is to explain the biological phenomena that stand behind such an excellent behavior of the contextual approach in all of these examples. Answering this question could help to discover a better alignment algorithm that profits from contextual information. Probably such an algorithm could be very efficient for the sequences belonging to some special class of sequences. The characterization of this class remains an interesting open problem.

CONCLUSIONS AND FURTHER WORKS

It is clear that the experimental analysis described in this work is just a beginning and cannot be treated as a definitive proof. Various improvements and other experiments can be envisaged.

Table 4. Summarized score for contextual versus noncontextual model. Score here corresponds to the frequency of properly aligned pairs of residues.

Reference	Protein families	Context	Noncontext	% of improvement
Ref 1	Short (< 25%)	0.5619	0.5260	6.83
	Short (20%–40%)	0.7323	0.7309	0.19
	Short (> 35%)	0.9004	0.8964	0.45
	Medium (< 25%)	0.4034	0.4091	−1.39
	Medium (20%–40%)	0.7951	0.7879	0.90
	Medium (> 35%)	0.9202	0.9198	0.04
	AVG	0.7379	0.7318	0.83
Ref 2	Short	0.6868	0.6633	3.52
	Medium	0.6580	0.6561	0.30
	AVG	0.6742	0.6602	2.12
Ref 3	Short	0.4008	0.4263	−5.49
	Medium	0.5880	0.5790	1.55
	AVG	0.4810	0.4917	−2.17
Ref 6	AVG	0.45	0.442	1.81
AVG	—	0.6674	0.6610	0.96

Table 5. The influence of orphans on the quality of the alignment.

Protein family	Context	Noncontext	% of improvement
Short	0.8918	0.8461	5.4
Medium	0.8593	0.8807	−2.4
AVG	0.8776	0.8613	1.89

Table 6. Families for which the contextual model gives much better alignments.

Protein family	No of sequences	Context	Noncontext	% of improvement
1ycc: cytochrome e	4	0.765	0.665	15.04
2trx: thioredoxin	4	0.671	0.468	43.38
1aboA: sh3	15	0.683	0.580	17.76
1uky: uridyl kin	24	0.541	0.464	20.91
sh3-2-ref6: sh3	6	0.553	0.454	21.81
sh3-3-ref6: sh3	5	0.430	0.214	100.93
AVG	—	0.606	0.474	29.11

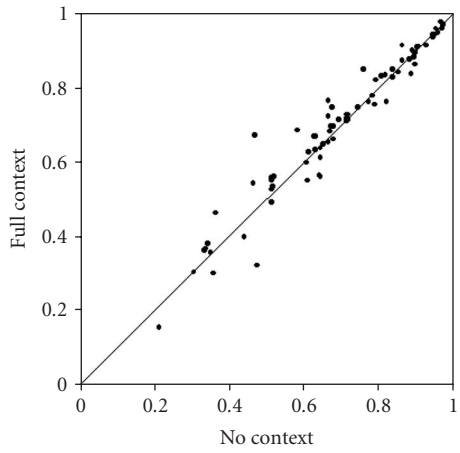
Wider and more distant contexts

In this paper we consider the simplest contextual model. The main idea of the context is to reflect the neighborhood of a given residue in the 3*D* protein structure. It suggests several possible extensions. One possibility is to consider a wider context, for example, two amino acids on each side. This approach is however limited by the huge size of substitution table ($20^6 = 64\,000\,000$ entries). The solution here is to consider a reduced context (ie, 20 amino acids can be divided into a small number of groups having similar biochemical properties (cf [5])).

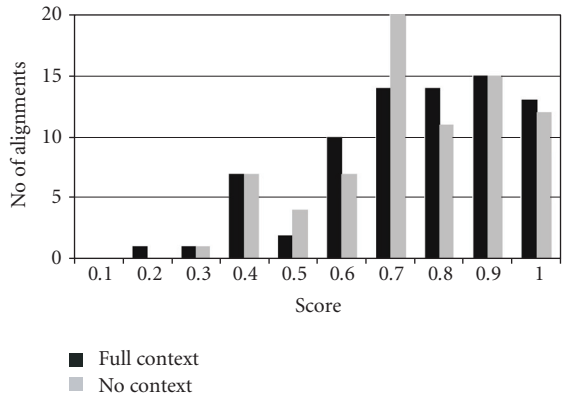
Another approach is to consider a more distant context. As an example look at the alignment

$$\ldots \text{CHCAD} \ldots$$
$$\ldots \text{HADGC} \ldots$$

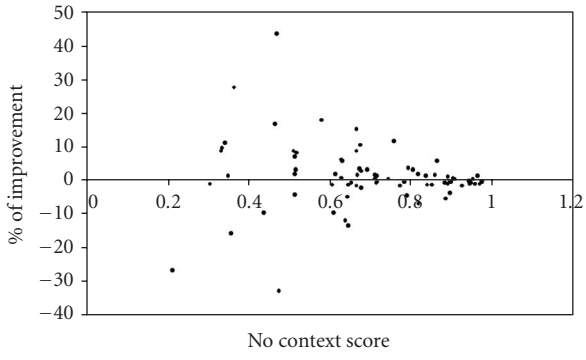
In the model we have presented, as a context we have taken two amino acids surrounding given substitution, that is, the left context of the substitution $C \mapsto D$ consists of amino acid H or A, and the right consists of A or G. It is however biologically motivated (by a secondary structure) to consider the contexts which are separated by



(a)



(b)



(c)

Figure 1. Comparison of Contextual and noncontextual scores. (a) SPS score comparison, (b) SPS score distributions, and (c) Improvement versus noncontextual score.

one position from the given residue, that is, the left context for the substitution $C \mapsto D$ is amino acid C or H, and the right context consists of D or C.

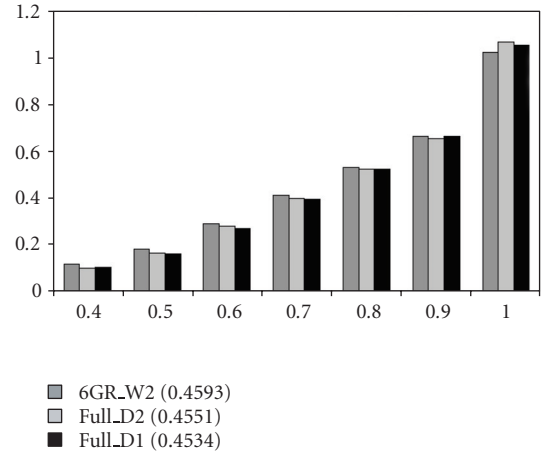


Figure 2. The entropy for substitution tables with a standard context (full_D1), a wider but grouped context (6GR_W2), and a more distant context (full_D2).

Preliminary results (see Figure 2) for the entropy of such defined substitution tables are very promising and encourage further research in this direction (more on entropy can be found in [8]).

Improvements for contextual multiple alignment algorithm

The algorithm presented in this paper follows the *progressive alignment* approach. The most popular algorithm and one of the most effective algorithms of this kind in the standard noncontextual model is ClustalW [9]. It contains a lot of additional heuristics. The challenging task is to design analogous improvements for the contextual model.

ACKNOWLEDGMENTS

This work was partially supported by the KBN Grant no 7 T11 F016 21. This work was also supported by the Open Society Institution (OSI) Grant no 18527.

REFERENCES

[1] Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK: Cambridge University Press; 1998.

[2] Feng DF, Doolittle RF. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* 1996;266:368–382.

[3] Gambin A, Lasota S, Szklarczyk R, Tiuryn J, Tyszkiewicz J. Contextual alignment of biological sequences. *Bioinformatics.* 2002; 18 (suppl 2): S116–27.

[4] Gambin A, Slonimski P. Hierarchical clustering based upon contextual alignment of proteins: a different way to approach phylogeny. *C R Biol.* 2005;328(1):11–22.
[5] Gambin A, Tyszkiewicz J. Substitution tables for contextual alignment. In: Proceedings of Journees Ouvertes Biologie Informatique Mathematique (JOBIM 2002) ; 2002 San Malo, France.
[6] Henikoff S, Henikoff JG, Pietrokovski S. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics.* 1999;15(6):471–479.
[7] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(2):10915–10919.
[8] Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 1991;219(3):555–565.
[9] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22(22):4673–4680.
[10] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4(4):406–425.
[11] Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 1999;27(13):2682–2690.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 132–138 • DOI: 10.1155/JBB.2005.132

RESEARCH ARTICLE

Selecting Genes by Test Statistics

Dechang Chen,[1] Zhenqiu Liu,[2] Xiaobin Ma,[3] and Dong Hua[4]

[1]*Division of Epidemiology and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA*
[2]*Bioinformatics Cell, TATRC, 110 North Market Street, Frederick, MD 21703, USA*
[3]*Department of Computer Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA*
[4]*Department of Computer Science, The George Washington University, 801 22nd St. NW, Washington, DC 20052, USA*

Received 28 April 2004; revised 22 November 2004; accepted 23 November 2004

Gene selection is an important issue in analyzing multiclass microarray data. Among many proposed selection methods, the traditional ANOVA F test statistic has been employed to identify informative genes for both class prediction (classification) and discovery problems. However, the F test statistic assumes an equal variance. This assumption may not be realistic for gene expression data. This paper explores other alternative test statistics which can handle heterogeneity of the variances. We study five such test statistics, which include Brown-Forsythe test statistic and Welch test statistic. Their performance is evaluated and compared with that of F statistic over different classification methods applied to publicly available microarray datasets.

INTRODUCTION

Microarrays provide information about the expression level of the genes represented on the array. Such gene expression profiling has been successfully applied to class prediction, where the purpose is to classify and predict the diagnostic category of a sample by its gene expression profile [1, 2, 3, 4]. Various machine learning methods are currently used for class prediction. However, the task of prediction by microarrays is challenging, due to a large number of genes (features) and a small number of samples involved in the problem. As a consequence, one has to identify a small subset of informative genes contributing most to the classification task. Performing feature selection is essential for microarray prediction problems, since high-dimensional problems usually involve higher computational complexity and bigger prediction errors.

Many methods have been proposed to select informative genes. One category of such work depends on the traditional t test statistic [5, 6, 7] and analysis of variance (ANOVA) F test statistic [8, 9]. While t is used for two-class prediction problems, F is used for multiclass problems. The statistics t and F are not only used in class prediction, they also apply to the class discovery [10, 11]. The main goal of class discovery is to identify subtypes of diseases. The major difference between class prediction and class discovery is that the former uses labeled samples while the latter uses unlabeled samples.

Although t and F have been commonly used in the analysis of gene expression data, there exists a misunderstanding on the roles of t and F. The test statistic t is used to detect the difference between the means of two populations and it has two versions depending on whether or not the two variances of the two populations are equal. The test statistic F is often used to detect the difference among the means of three or more populations under the assumption that the variances of the involved populations are equal. Of course, the F statistic can be used to detect the difference between the means of two populations. In doing this, one can show that the F statistic is equivalent to the t statistic based on the equal variance, that is, one procedure rejects the null hypothesis that the two populations have the same mean if and only if the other procedure rejects the null hypothesis. In analyzing gene expression data, the t statistic is based on unequal variances so that its extension will never reach the ANOVA F. Therefore, for multiclass prediction problems, it is natural to explore other statistics which do not assume equal variances.

Correspondence and reprint requests to Dechang Chen, Division of Epidemiology and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA, E-mail: dchen@usuhs.mil

In this paper, we study the effect on multiclass prediction results of gene selection from six test statistics: ANOVA F test statistic, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, Cochran test statistic, and Kruskal-Wallis test statistic. The five last test statistics can be viewed as extensions of the t statistic used in two-class prediction problems. Their performance will be compared with that of the F statistic.

This paper is organized as follows. In "models and methods," we describe the statistical model for gene expression levels, test statistics, and our method to select genes. In "experimental results," we investigate the effect of test statistics on the classification results by using our gene selection approach and different machine learning techniques, applied to five publicly available microarray datasets. Our conclusion is given in "conclusion."

MODELS AND METHODS

In this section, we will first introduce a general statistical model for gene expression values and describe test statistics for testing the equality of the class means. We then present our approach to select genes using power and correlation.

Statistical model

Assume there are k (≥ 2) distinct tumor tissue classes for the problem under consideration and there are p genes (inputs) and n tumor mRNA samples (observations). Suppose X_{gs} is the measurement of the expression level of gene g from sample s for $g = 1, \ldots, p$ and $s = 1, \ldots, n$. In terms of an expression matrix $\mathbf{G}$, we may write

$$\mathbf{G} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix}. \tag{1}$$

It is seen that the columns and rows of the expression matrix $\mathbf{G}$ correspond to samples and genes, respectively. Note that $\mathbf{G}$ is a matrix consisting of data highly processed through preprocessing techniques that include image analysis and normalization and often logarithmic transformations. We assume that the data $\mathbf{G}$ are standardized so that the genes have mean 0 and variance 1 across samples. Given a fixed gene, let Y_{ij} be the expression level from the jth sample of the ith class. Note that these Y_{ij} come from the corresponding row of $\mathbf{G}$. For example, for gene 1, Y_{ij} are a rearrangement of the first row of $\mathbf{G}$. We consider the following general model for Y_{ij}:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \text{for } i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i \tag{2}$$

with $n_1 + n_2 + \cdots + n_k = n$. In the model, μ_i is a parameter representing the mean expression level of the gene in class i, ϵ_{ij} are the error terms such that ϵ_{ij} are independent normal random variables, and

$$E(\epsilon_{ij}) = 0, \quad V(\epsilon_{ij}) = \sigma_i^2 < \infty, \tag{3}$$

for $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$. Schematically, the expression levels Y_{ij} look like the following:

Classes				
1	2	3	$\cdots$	k
Y_{11}	Y_{21}	Y_{31}	$\cdots$	Y_{k1}
Y_{12}	Y_{22}	Y_{32}	$\cdots$	Y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\cdots$	Y_{2n_2}	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	Y_{3n_3}	$\cdots$	$\cdots$
Y_{1n_1}	$\cdots$	$\cdots$	$\cdots$	Y_{kn_k}

Note that if the variances are equal, that is, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, then the above model is simply the commonly used one-way ANOVA model. For the microarray data, we believe that heterogeneity in the variances is more realistic, since different σ_i may describe different variations of the gene expression across classes.

One of the main tasks associated with the above model is to detect whether or not there is some difference among the means $\mu_1, \mu_2, \ldots, \mu_k$. For the case of homogeneity of variances, the well-known ANOVA F test is the optimal test to accomplish the task [12, 13]. However, with heterogeneity of the variances, the task is challenging and is closely related to the well-known Behrens-Fisher problem [14]. When the sample sizes in all classes are equal, that is, $n_1 = n_2 = \cdots = n_k$, the presence of heterogeneous variances of the errors only slightly affects the F test. When the sample sizes are unequal, the effect is serious [15]. The actual type-I error is inflated if smaller sizes n_i are associated with larger variances σ_i^2. In addition, the significance levels are smaller than anticipated if larger sizes n_i are associated with larger variances σ_i^2. The above indicates that for our model, the F test may not be appropriate for testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus H_1: not all the μ_i are equal. Therefore some alternatives to the F test are worthy of investigating.

Test statistics

After introducing the statistical model for gene expression values, we now turn to the test statistics used to test the equality of the class means for a fixed gene. We will consider the following six test statistics. The first five are parametric test statistics, while the last one is nonparametric.

(a) *ANOVA F test statistic*. The definition of this test is

$$\mathrm{F} = \frac{(n-k)\sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{(k-1)\sum (n_i - 1)s_i^2}, \tag{4}$$

where $\bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{k} n_i \bar{Y}_{i\cdot}/n$, and $s_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2/(n_i - 1)$. For simplicity, we use $\sum$ to indicate the sum is taken over the index i. Under H_0 and assuming variance homogeneity, this well-known test statistic has a distribution of $F_{k-1,n-k}$ [13].

(b) *Brown-Forsythe test statistic* [16]. This is given by

$$\mathrm{B} = \frac{\sum n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sum (1 - n_i/n) s_i^2}. \tag{5}$$

Under H_0, B is distributed approximately as $F_{k-1,\nu}$, where

$$\nu = \frac{\left[\sum (1 - n_i/n) s_i^2\right]^2}{\sum (1 - n_i/n)^2 s_i^4/(n_i - 1)}. \tag{6}$$

(c) *Welch test statistic* [17]. This is defined as

$$\mathrm{W} = \frac{\sum w_i (\bar{Y}_{i\cdot} - \sum h_i \bar{Y}_{i\cdot})^2}{(k-1) + 2(k-2)(k+1)^{-1} \sum (n_i - 1)^{-1} (1 - h_i)^2} \tag{7}$$

with $w_i = n_i/s_i^2$ and $h_i = w_i/\sum w_i$. Under H_0, W has an approximate distribution of F_{k-1,ν_w}, where

$$\nu_w = \frac{k^2 - 1}{3 \sum (n_i - 1)^{-1} (1 - h_i)^2}. \tag{8}$$

(d) *Adjusted Welch test statistic* [18]. It is similar to the Welch test statistic and defined to be

$$\mathrm{W}^* = \frac{\sum w_i^* (\bar{Y}_{i\cdot} - \sum h_i^* \bar{Y}_{i\cdot})^2}{(k-1) + 2(k-2)(k+1)^{-1} \sum (n_i - 1)^{-1} (1 - h_i^*)^2}, \tag{9}$$

where $w_i^* = n_i/(\phi_i s_i^2)$ with ϕ_i chosen such that $1 \le \phi_i \le (n_i - 1)/(n_i - 3)$, and $h_i^* = w_i^*/\sum w_i^*$. Under H_0, W* has an approximate distribution of F_{k-1,ν_w^*}, where

$$\nu_w^* = \frac{k^2 - 1}{3 \sum (n_i - 1)^{-1} (1 - h_i^*)^2}. \tag{10}$$

In this paper, we choose $\phi_i = (n_i + 2)/(n_i + 1)$, since this choice provides reliable results for small sample sizes n_i and a large number (k) of populations [18].

(e) *Cochran test statistic* [19]. This test statistic is simply the quantity appearing in the numerator of the Welch test statistic W, that is,

$$\mathrm{C} = \sum w_i (\bar{Y}_{i\cdot} - \sum h_i \bar{Y}_{i\cdot})^2, \tag{11}$$

where w_i and h_i are given in (c). Under H_0, C has an approximate distribution of χ^2_{k-1}.

(f) *Kruskal-Wallis test statistic.* This is the well-known nonparametric test and is given by

$$\mathrm{H} = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1), \tag{12}$$

where R_i is the rank sum for the ith class. The ranks assigned to Y_{ij} are those obtained from ranking the entire set of Y_{ij} (use the average rank in case of tied values). Assuming each $n_i \ge 5$, then under H_0, H has an approximate distribution of χ^2_{k-1} [20].

Gene selection

With the test statistics introduced above, we are able to discuss the issue of gene selection. It has been well demonstrated in the literature that gene selection is an important issue in microarray data analysis. It is also known that with a large number of genes (usually in thousands) present, no practical method is available to locate the best set of genes, that is, the smallest subset of genes that offer optimal prediction accuracy. In this paper, the focus lies in comparing the performance of different test statistics in selecting genes for the classification of tumors based on gene expression profiles. Identifying a gene selection process to achieve good classification results is not the purpose of this paper. To make the comparison straightforward, we adopt the simplest gene selection approach as follows. First, we formulate the expression levels of a given gene by a one-way ANOVA model, as shown in "statistical model." We then use the test statistics in "test statistics" to determine the power of genes in discriminating between tumor types. Given a test statistic $\mathcal{F}$, we define the *discrimination power* of a gene as the value of $\mathcal{F}$ evaluated at the n expression levels of the gene. This definition is based on the fact that with larger $\mathcal{F}$ the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ will be more likely rejected. Therefore, the higher the discrimination power is, the more powerful the gene is in discriminating between tumor types. Finally, we choose as informative genes those genes having high power of discrimination.

We note that the discrimination power of genes could be determined equally well by the p value from $\mathcal{F}$. However, due to small sizes n_i, it is hard to justify the approximation of the known distribution to $\mathcal{F}$. Therefore the p values may not reflect the actual functionality of $\mathcal{F}$. This drawback is overcome by using the value of $\mathcal{F}$ to determine the power of discrimination. Another obvious benefit is that using the value of $\mathcal{F}$ will greatly simplify the calculation.

In [18], extensive simulations have been conducted to examine the behavior of some test statistics for testing the equality of population means. The test statistics studied include B, W, W*, F, and C. The results show that with homogeneity of the variances, the ANOVA F test is the optimal test, as stated in "statistical model." However, this assumption of homogeneity is rarely met in practice. Under heterogeneity of variances, the simulation results in [18] show that the test statistics B, W, and W* provide acceptable control of type I errors. This implies that the genes identified by B, W, and W* are more likely to be powerful than those by F and C in discriminating between tumor types, and thus the prediction errors resulting from B, W, and W* are expected to be lower than those from F and C. The nonparametric test statistic H can be applied to data with less restriction, for example, ordinal data, and thus is expected to perform worse than test statistics such as B, W, W*, and C. The above discussion will be further verified by our experiments on gene expression data conducted in "experimental results."

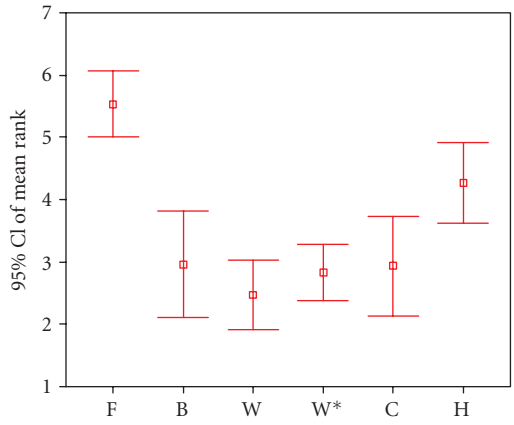


Figure 1. Relative performances of test statistics based on the average errors.

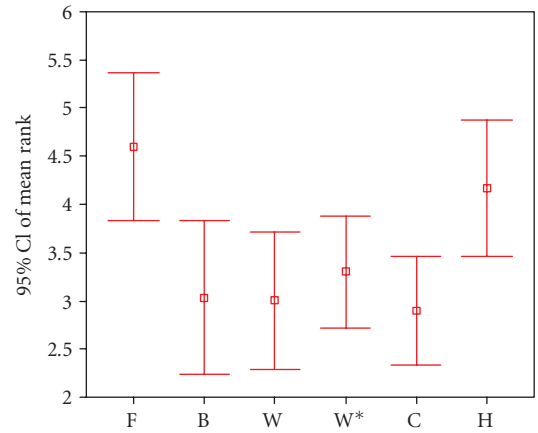


Figure 2. Relative performances of test statistics based on the median errors.

EXPERIMENTAL RESULTS

In this section we investigate the effect on gene selection of the six test statistics introduced in "test statistics." Five gene expression datasets and five prediction methods are used for this purpose. The performances of the test statistics are evaluated in terms of class prediction errors.

Datasets

We considered five multiclass gene expression datasets: leukemia72 [1], ovarian [21], NCI [22, 23], lung cancer [24], and lymphoma [25]. Table 1 presents more details of the datasets.

Comparison of test statistics

The gene selection procedure described above depends on the test statistics. Given a gene selection process from a test statistic, different classification methods may lead to different prediction errors. In our experiments, we used the following five prediction methods: naive Bayes, nearest neighbor, linear perceptron, multilayer perceptron neural network with 5 nodes in the middle layer, and support vector machines with a second-order polynomial kernel. All the algorithms are from Matlab PRTools 3.01 by Robert P. W. Duin.

To calculate the overall prediction error, we used leave one out (LOO) cross-validation. For a dataset with n samples, this method involves n separate runs. For each of the runs, $n-1$ data points are used to train the model and then prediction is performed on the remaining data point. The overall prediction error is the sum of the errors on all n runs.

Table 2 presents a comparison of the six test statistics when 50 informative genes were used. In the table, F, B, W, W*, C, and H represent the ANOVA F test statistic, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, Cochran test statistic, and Kruskal-Wallis test statistic, respectively. The first number in each cell denotes the average of 5 prediction errors from 5 different classification methods. The second number in each cell is the median of the 5 prediction errors. The results in the table suggest that B, W, W*, and, C perform better than F and H. Similar to Table 2, Tables 3 and 4 present comparison results with 100 and 200 informative genes, respectively.

Results in Tables 2, 3, and 4 may be summarized in a way by figures. Consider the average errors in the tables. For a fixed dataset and fixed number of informative genes, the performances of the six test statistics can be ranked. The fifteen ranks achieved by a test statistic could be used to obtain a 95% confidence interval of the mean rank for the test statistic. The corresponding bar chart plotting six confidence intervals is given in Figure 1. The bar chart based on the median errors in Tables 2, 3, and 4 is presented in Figure 2. Clearly, both figures show that B, W, W*, and C outperform F and H. These results indicate that the proposed models in "statistical model" without assuming equal variances are preferred to those assuming equal variances.

We note that in the above experiments, the performance of C is comparable to those of B, W, and W*. This does not look consistent with the discussion in "gene selection." One reason might be that we only examined 5 datasets in this paper. Our opinion is that if more data sets are explored, the overall performance of C should be worse than that of B, W, or W*. We leave this as our future work.

Before concluding, we point out that it is useful to assess the importance of genes selected by the test statistics from the biological perspective. Since this is not the focus of our research work in this paper, below we only provide a simple example to examine some genes selected by the Brown-Forsythe test statistic for the leukemia dataset. This dataset was also studied by Getz et al [26]. They extracted the stable clusters of genes by the coupled two-way clustering analysis and concluded that those genes grouped into the same cluster share certain biological significance such as on the same pathway. Among the top 50

Table 1. Multiclass gene expression datasets.

Dataset	Leukemia72	Ovarian	NCI	Lung cancer	Lymphoma
No of genes	6817	7129	9703	918	4026
No of samples	72	39	60	73	96
No of classes	3	3	9	7	9

Table 2. Performances of the test statistics with 50 informative genes.

Dataset	F	B	W	W*	C	H
Leukemia	3.4	2.4	2.8	2.8	3.2	3.0
	3	2	3	3	3	3
Ovarian	0.2	0.0	0.0	0.0	0.0	0.0
	0	0	0	0	0	0
NCI	36.0	32.0	27.4	26.0	27.0	35.4
	35	29	27	27	27	35
Lung cancer	17.6	17.0	17.6	17.6	18.0	18.0
	17	17	18	18	18	18
Lymphoma	23.8	19.8	14.0	14.0	12.8	22.0
	23	19	12	12	13	20

Table 3. Performances of the test statistics with 100 informative genes.

Dataset	F	B	W	W*	C	H
Leukemia	3.4	3.0	3.0	3.0	3.2	3.0
	3	3	4	3	3	3
Ovarian	0.2	0.0	0.0	0.0	0.0	0.0
	0	0	0	0	0	0
NCI	33.0	22.6	23.8	25.2	25.2	31.6
	33	22	25	26	26	31
Lung cancer	12.2	12.2	11.4	12.2	12.2	15.8
	12	12	11	11	11	14
Lymphoma	21.8	19.2	13.0	13.8	14.4	18.2
	17	16	12	12	12	18

Table 4. Performances of the test statistics with 200 informative genes.

Dataset	F	B	W	W*	C	H
Leukemia	3.0	3.0	2.4	2.8	1.8	2.4
	3	3	2	3	1	2
Ovarian	0.4	0.2	0.2	0.2	0.2	0.4
	0	0	0	0	0	0
NCI	25.6	22.6	22.6	22.8	22.2	25.6
	26	22	24	25	24	25
Lung cancer	15.2	12.6	14.2	13.2	12.8	13.2
	13	11	12	12	12	11
Lymphoma	21.2	18.8	12.0	12.6	12.8	16.2
	15	14	8	9	8	14

Table 5. Mapping from genes selected by the Brown-Forsythe test statistic for the leukemia data to clusters of genes of interest provided by Getz et al [26].

Gene description	Access number	Cluster by Getz et al [26]	B
GB DEF = T-cell antigen receptor gene T3-delta	X03934	LG5	70.808014
Protein tyrosine kinase related mRNA sequence	L05148	LG5	43.676056
CD33CD33 antigen (differentiation antigen)	M23197	LG1	42.435883
GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA	U23852 s	LG5	35.120228
T-cell surface glycoprotein CD3 epsilon chain precursor	M23323 s	LG5	35.028965
CTSD (cathepsin D) (lysosomal aspartyl protease)	M63138	LG1	34.865067
HLA class II histocompatibility antigen, DR alpha chain precursor	X00274	LG6	31.882597
HLA class I histocompatibility antigen, F alpha chain precursor	X17093	LG6	31.83585
Leukotriene C4 synthase (LTC4S) gene	U50136 rna1	LG1	31.183104
RNS2 (ribonuclease 2) (eosinophil-derived neurotoxin (EDN))	X16546	LG1	29.52516
TIMP2 (tissue inhibitor of metalloproteinase 2)	M32304 s	LG1	28.233025
LMP2 gene extracted from *Homo sapiens* genes TAP1, TAP2, LMP2, LMP7, and DOB	X66401 cds1	LG6	27.11849

informative genes from the Brown-Forsythe test statistic, 12 were mapped to the clusters of genes of interest given in [26]. Table 5 shows the information about the gene names, access numbers, corresponding clusters as well as the values of the Brown-Forsythe statistic. For details on the explanation of biological significance of clusters LG1, LG5, and LG6, readers are referred to [26].

CONCLUSION

In this paper, we have compared the performance of different test statistics in selecting genes for multiclassification of tumors using gene expression data. Experiments show (a) the model for gene expression values without assuming equal variances is more appropriate than that assuming equal variances; (b) Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, and Cochran test statistic perform much better than ANOVA F test statistic and Kruskal-Wallis test statistic.

DISCLAIMER

The opinions expressed herein are those of the authors and do not necessarily represent those of the Uniformed Services University of the Health Sciences and the Department of Defense.

ACKNOWLEDGMENTS

The authors thank Dr. Hanchuan Peng of Lawrence Berkeley National Laboratory for providing the NCI, lung cancer, and lymphoma data. The authors also thank the referees for providing many valuable comments. D. Chen was supported by the National Science Foundation grant CCR-0311252.

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531–537.

[2] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA.* 2001;98(26):15149–15154.

[3] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Amer Statist Assoc.* 2002;97(457):77–87.

[4] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA.* 2002;99(10):6567–6572.

[5] Xiong M, Jin L, Li W, Boerwinkle E. Computational

methods for gene expression-based tumor classification. *Biotechniques.* 2000;29(6):1264–1270.
[6] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics.* 2002;18(1):39–50.
[7] Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform Ser Workshop Genome Inform.* 2002;13:51–60.
[8] Ghosh D. Singular value decomposition regression models for classification of tumors from microarray experiments. In: *Proceedings of the 2002 Pacific Symposium on Biocomputing.* Lihue, Hawaii: 2002:18–29.
[9] Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics.* 2002;18(9):1216–1226.
[10] Ding C. Analysis of gene expression profiles: class discovery and leaf ordering. In: *Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2002).* Washington, DC: 2002:127–136.
[11] Li W, Fan M, Xiong M. SamCluster: An integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics.* 2003;19(7):811–817.
[12] Lehman EL. *Testing Statistical Hypotheses.* 2nd ed. NY: Wiley; 1986.
[13] Neter J, Kutner MH, Nachtsheim CJ, et al. *Applied Linear Statistical Models.* 4th ed. Chicago, Ill: McGraw-Hill; 1996.
[14] Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics. Volume 2A: Classical Inference and the Linear Model.* 6th ed. London: Edward Arnold; 1999.
[15] Montgomery DC. *Design and Analysis of Experiments.* 5th ed. NY: Wiley; 2001.
[16] Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. *Technometrics.* 1974;16:129–132.
[17] Welch BL. On the comparison of several mean values: An alternative approach. *Biometrika.* 1951;38:330–336.
[18] Hartung J, Argaç D, Makambi KH. Small sample properties of tests on homogeneity in one-way ANOVA and meta-analysis. *Statist Papers.* 2002;43:197–235.
[19] Cochran WG. Problems arising in the analysis of a series of similar experiments. *J R Stat Soc Ser C Appl Stat.* 1937;4:102–118.
[20] Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences.* 7th ed. NY: Wiley; 1999.
[21] Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci USA.* 2001;98(3):1176–1181.
[22] Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* 2000;24(3):227–235.
[23] Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet.* 2000;24(3):236–244.
[24] Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA.* 2001;98(24):13784–13789.
[25] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000;403(6769):503–511.
[26] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA.* 2000;97(22):12079–12084.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 139–146 • DOI: 10.1155/JBB.2005.139

RESEARCH ARTICLE

Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences

Jianbo Gao,[1] Yan Qi,[2] Yinhe Cao,[3] and Wen-wen Tung[4]

[1]*Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL 32611-6200, USA*
[2]*Department of Biomedical Engineering, Whitaker Institute, Johns Hopkins University, Baltimore, MD 21205, USA*
[3]*BioSieve, 1026 Springfield Drive, Campbell, CA 95008, USA*
[4]*National Center for Atmospheric Research, Boulder, CO 80307-3000, USA*

Received 24 May 2004; revised 30 August 2004; accepted 3 September 2004

Most codon indices used today are based on highly biased nonrandom usage of codons in coding regions. The background of a coding or noncoding DNA sequence, however, is fairly random, and can be characterized as a random fractal. When a gene-finding algorithm incorporates multiple sources of information about coding regions, it becomes more successful. It is thus highly desirable to develop new and efficient codon indices by simultaneously characterizing the fractal and periodic features of a DNA sequence. In this paper, we describe a novel way of achieving this goal. The efficiency of the new codon index is evaluated by studying all of the 16 yeast chromosomes. In particular, we show that the method automatically and correctly identifies which of the three reading frames is the one that contains a gene.

INTRODUCTION

Gene identification is one of the most important tasks in the study of genomes. In order to be successful, a gene-finding algorithm has to incorporate good indices for the protein coding regions. In the past two decades, a number of useful codon indices have been proposed. They include the codon bias index (CBI) (Bennetzen and Hall [1]), the codon adaptation index (CAI) (Sharp and Li [2]; Jansen et al [3]), the YZ score (Zhang and Wang [4]), measures based on differences in codon usage (Staden and McLachlan [5]), hexamer counts (Claverie and Bougueleret [6]; Farber et al [7]; Fickett and Tung [8]), codon position asymmetry (Fickett [9]), autocorrelations and nucleotide frequencies (Shulman et al [10]; Fickett [9]; Borodovsky et al [11]), entropy (Almagor [12]), and periodicities, especially the period-3 feature of a nucleotide sequence in the coding regions (Fickett [9]; Silverman and Linsker [13]; Chechetkin and Turygin [14]; Tiwari et al [15]; Trifonov [16]; Yan et al [17]; Anastassiou [18]; Issac et al [19]; Kotlar and Lavner [20]). Most of them mainly capture the feature of highly biased nonrandom usage of codons in the coding regions. The background of a DNA sequence, be it a coding or noncoding sequence, however, is fairly random. Consequentially, a DNA sequence can be characterized as a random fractal. Is it possible to develop a new codon index by simultaneously incorporating the fractal and periodic features of a DNA sequence? The aim of this paper is to develop a simple method to achieve such a goal. Since the codon index obtained this way complements existing codon indices, it has the potential of being incorporated into existing gene identification algorithms so that the accuracy of those algorithms can be improved and their training be simplified.

The novel codon index proposed here is based on two incompatible features of DNA sequences: the period-3 and the fractal features. It has been known for a while that a DNA sequence exhibits fractal properties, with noncoding regions often possessing, but coding regions often lacking long-range correlations (Li and Kaneko [21]; Peng et al [22]; Voss [23]). Roughly speaking, a fractal means a part is similar to another part or to the whole,

Correspondence and reprint requests to Jianbo Gao, Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL 32611-6200, USA, E-mail: gao@ece.ufl.edu

and hence, does not possess any well-defined scale (Mandelbrot [24]). However, it has been recognized that the biased nonrandom use of codons in coding regions often defines a period-3 feature in the coding regions. Period-3 is a specific scale and hence, is incompatible with the concept of fractal. We show here that a convenient codon index can be developed by exploiting this incompatibility. This is achieved by quantifying the deviation of a DNA sequence from its fractal behavior due to the period-3 feature. Amazingly, this simple measure not only correlates well with coding regions, but also automatically and correctly identifies which of the three reading frames is the correct one (ie, containing a gene). In this paper, we will illustrate the idea and evaluate the proposed index by studying all of the 16 yeast chromosomes.

DATABASES AND METHODS

Database

The yeast chromosome sequences and the associated annotation data used for the analysis are based on sequence dated 1 October 2003 in the *Saccharomyces* Genome Database (http://www.yeastgenome.org) and can be obtained from ftp://genome-ftp.stanford.edu/pub/yeast/data_download.

The period-3 feature

The period-3 feature of a DNA sequence has been used to develop important codon indices (Fickett [9]; Silverman and Linsker [13]; Chechetkin and Turygin [14]; Tiwari et al [15]; Trifonov [16]; Yan et al [17]; Anastassiou [18]; Issac et al [19]; Kotlar and Lavner [20]). The existence of this feature can be shown, for example, by Fourier spectral analysis. In order to apply Fourier transform, first one has to obtain one or more numerical sequences from a DNA sequence. A common mapping scheme is to construct four binary sequences from a DNA sequence, one for each base. For instance, when nucleotide base "A" is concerned, a sequence $u(n)$ is assigned 1's at those positions where "A" is present, and 0's otherwise. Take sequence

$$\mathrm{S} = \mathrm{AATCGGCCCCGAT} \tag{1}$$

as an example. One obtains the following binary sequences:

$$\begin{aligned} u(\mathrm{A}) &= 1100000000010, \\ u(\mathrm{C}) &= 0001001111000, \\ u(\mathrm{G}) &= 0000110000100, \\ u(\mathrm{T}) &= 0010000000001. \end{aligned} \tag{2}$$

Such a scheme has been used, for example, by Voss [23] and Kotlar and Lavner [20]. Alternatively, one can obtain numerical sequences using the following mapping rules. (a) C or G $\rightarrow$ $u(n) = +1$; A or T $\rightarrow$ $u(n) = -1$. This rule suggested by Azbel [25] maps a DNA sequence into a sequence of weak/strong hydrogen bonds. (b) C or T$\rightarrow$ $u(n) = +1$; A or G$\rightarrow$ $u(n) = -1$. This scheme was proposed by Peng et al [22] and maps a DNA sequence into a sequence of purine/pyrimidine. When a numerical sequence $u(n)$ is obtained, the discrete fourier transform (DFT) can be used to compute its spectrum $U(k)$, which is given by

$$U(k) = \sum_{n=0}^{N-1} u(n) e^{(-2\pi/N)nk}, \quad 0 \le k \le N-1, \tag{3}$$

where N is the length of $u(n)$ and k corresponds to the discrete frequency of $(2\pi/N)k$ or a period of (N/k). $U(k)$ can be conveniently used to identify characteristic periodicities of $u(n)$. Since a coding DNA sequence is comprised of codons (units of three nucleotide bases) and the nucleotide usage in a coding sequence is highly biased and nonrandom, a period of 3 is often present in the coding sequence $u(n)$. This feature is usually referred to as "period-3." Consequently, the DFT magnitude or power spectrum density of $u(n)$ often displays a distinct peak at $k = N/3$ (or at a frequency around $[N/3]$ when $N/3$ is not an integer). However, the period-3 feature is usually lacking or weak in noncoding regions (Fickett [9]; Silverman and Linsker [13]; Chechetkin and Turygin [14]; Tiwari et al [15]; Trifonov [16]; Yan et al [17]; Anastassiou [18]; Issac et al [19]; Kotlar and Lavner [20]). To illustrate this idea, we use Peng's mapping rule to construct $u(n)$ from the coding/noncoding DNA sequences of yeast and perform DFT on $u(n)$ (for simplicity, we have chosen $N = 1026$). Typical DFT magnitudes $|U(k)|$ for coding and noncoding regions are shown in Figure 1. A strong peak is observed for $|U(k)|$ at $k = 1026/3$ of the coding region while no such feature is observed for the noncoding region.

Fractal property and the DFA technique

For other analyses, especially fractal analysis, it is more handy to construct a random walk (called DNA walk) from the DNA sequence. The walk $y(n)$ is generated by forming a partial sum of the $u(i)$ sequence constructed from the DNA sequence

$$y(n) = \sum_{i=1}^{n} u(i), \quad n = 1, 2, 3 \ldots. \tag{4}$$

Several different versions of DNA walks have been proposed based on different mapping rules for $u(n)$. The recently proposed 3D DNA walk is also called Z curve (Yan et al [17]; Zhang et al [26]). Note that the DNA walks based on the two 1D mapping rules mentioned above (Azbel [25]; Peng et al [22]) are equivalent to the z-component and x-component of the Z curve, respectively. Other types of multidimensional DNA walks have also been suggested (Berthelsen et al [27]; Cebrat et al [28]). In this work, we will employ the x-component of the Z curve (A or T $\rightarrow$ $u(n) = +1$; C or G $\rightarrow$ $u(n) = -1$) for further analysis because of its simplicity and efficiency (Stanley et al [29]). An example is shown in Figure 2 for the first

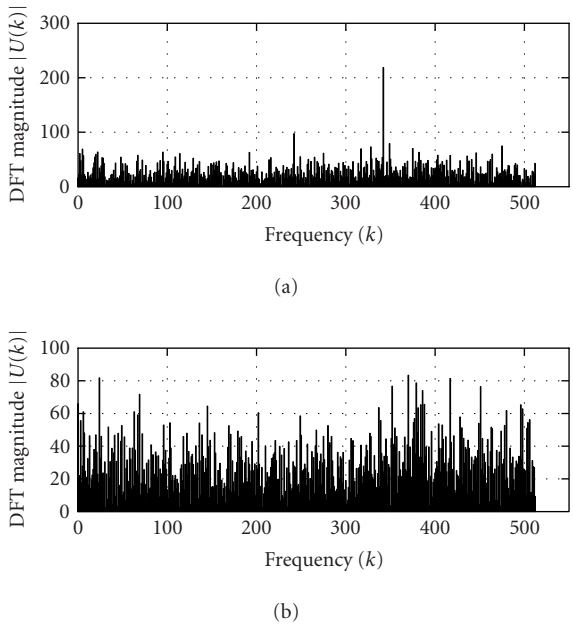


FIGURE 1. Representative DFT magnitudes for (a) coding and (b) noncoding regions in yeast chromosome I.

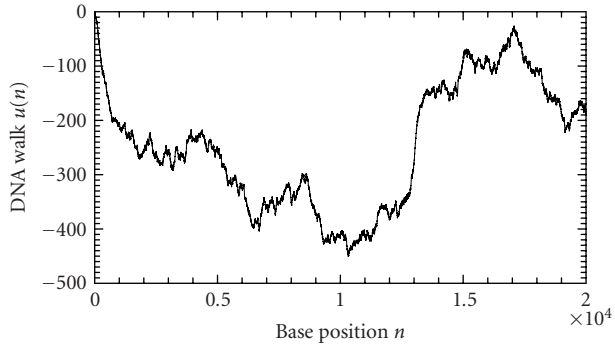


FIGURE 2. An example of a DNA walk constructed from the first 20 000 bases of the chromosome I of yeast.

20 000 bases of the chromosome I of yeast. Note that such a mapping generates an equivalent (except differing by a sign) DNA walk for the reverse strand of a DNA sequence. Hence, analysis based on such a mapping processes both strands of a DNA sequence simultaneously.

While DNA walks are useful in many applications, interpretation of some computational results, such as long-range correlations (Stanley et al [29]), are sometimes problematic, due to patchiness effects along a DNA sequence (Karlin and Brendel [30]). To remove such patchiness effects, a method called detrended fluctuation analysis (DFA) was developed by Peng et al [31] and has been used to identify characteristic patch sizes (Viswanathan et al [32]). DFA works as follows: first divide a given DNA walk of length N into $\lfloor N/l \rfloor$ nonoverlapping segments (where the notation $\lfloor x \rfloor$ denotes the largest integer that is not greater than x), each containing l nucleotides; then define the local trend in each segment to be the ordinate of a linear least-squares fit for the DNA walk in that segment; finally compute the "detrended walk," denoted by $y_l(n)$, as the difference between the original walk $y(n)$ and the local trend. The following scaling behavior (ie, fractal property) has been found for many DNA walks studied:

$$[F_d(l)]^2 = \left\langle \sum_{i=1}^{l} y_l(i)^2 \right\rangle \propto l^{2H}, \tag{5}$$

where the angle brackets denote ensemble average of all the segments and $F_d(l)$ is the average variance over all segments. The exponent H is often called the "Hurst parameter" (Mandelbrot [24]). When $H = 0.5$, the DNA walk is similar to a standard random walk. When $H > 0.5$, the DNA walk possesses long-range correlations. Statistically speaking, a noncoding region is often more likely to possess the long-range correlation properties (Stanley et al [29]). This feature, together with the DFA technique, was used by Ossadnik et al [33] to develop a coding sequence finder for genomes with long noncoding regions. To further illustrate the ideas, we analyze the coding and noncoding sequences of the yeast genome using the DFA technique. For a coding/noncoding sequence of length N, first a DNA walk $y(n)$ is constructed according to Peng's mapping rule. Then the detrended fluctuation $F_d(l)$ is computed according to (5) for a series of segment sizes l ($l < N$). In practice, l is often chosen to be the power of a common base r, that is, $l(j) = r^j$, $j = 1, 2, \ldots, \log_r^N$. Notice that $\log F_d(l) \sim H \log l$, $F_d(l)$ is approximately linear on double logarithmic scale when l is within a certain range $[l_0, l_1]$. A linear least-squares fit of data in this range produces a straight line with slope a and intersect b from which we can get an estimate of the Hurst parameter $H = a/2$. Figure 3 shows a log-log plot of $F_d(l)$ versus l for (a) a coding and (b) a noncoding sequence of yeast chromosome I. The two sequences are of lengths (a) $N = 1742$ and (b) $N = 3598$. We choose l to increment with the base $r = 2$ (also for all following analysis that concerns DFA) and the fitting range with the best scaling property is found to be $[l_0, l_1] = [2^2, 2^8]$ for both (a) and (b). Within this range, the Hurst parameters are (a) $H = 0.54$ and (b) $H = 0.62$. The nice scaling law in $[l_0, l_1]$ indicates that DNA sequences are fractals. Often, the Hurst parameters in noncoding regions are larger than those in coding regions, suggesting that noncoding regions often possess stronger long-range correlations. Sometimes this feature is termed lesser complexity in noncoding regions (Ossadnik et al [33]; Stanley et al [29]).

Deviation from fractal scaling due to period-3 signal

Intuitively, when a periodicity exists in a sequence, the fractal scaling law does not hold at that particular "scale" defined by the periodicity. Specifically, for DNA

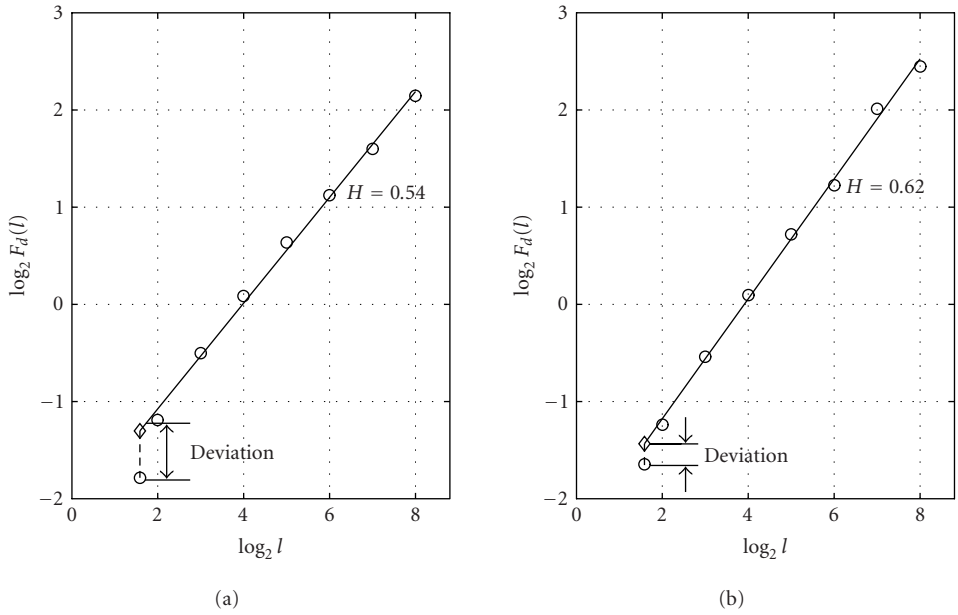


Figure 3. Representative period-3 fractal deviation (PFD) for (a) coding and (b) noncoding regions in yeast chromosome I.

sequences, a normally strong period-3 signal in coding regions causes a "deviation" from the sequence's fractal "background." On the contrary, for a noncoding region, such deviation, if any, is typically much smaller than that of a coding region. Based on the DFA technique, we have developed a novel codon index which quantifies such deviation from the fractal scaling law due to the period-3 feature, which we will denote by period-3 fractal deviation (PFD) for simplicity.

For a DNA walk $y(n)$ of length N, after computing its detrended fluctuation using DFA and identifying the best fitting range $[l_0, l_1]$, an approximation of $F_d(l)$ can be obtained by

$$\log \hat{F}_d(l) = a \log l + b. \tag{6}$$

The deviation of $y(n)$ is defined as the difference between $\log \hat{F}_d(l)$ and $\log F_d(l)$ at $l = 3$, that is,

$$\text{PFD} = \left| \log \hat{F}_d(3) - \log F_d(3) \right|. \tag{7}$$

To verify our intuition about the capacity of this index in distinguishing coding and noncoding regions, we have computed the PFD value for a large number of the verified open reading frames (ORFs) and noncoding segments from all of the 16 yeast chromosomes. An example of representative PFD values for coding and noncoding regions is shown in Figure 3. We observe that for the coding region, the fluctuation $F_d(l)$ at $l = 3$ deviates severely from the power-law relation (ie, the straight line in a log-log plot in Figure 3), while the deviation for the noncoding region is relatively small.

One may wonder if any DNA segment that belongs to a coding region has a large PFD. In fact, this is not the case. The quantification of the period-3 feature by the deviation from fractal scaling is reading-frame dependent. When the coding segment starts with the gene-containing reading frame (the first nucleotide of a codon), the period-3 feature collides with the DFA technique at the scale of $l = 3$ and results in a large PFD. When the segment starts with an incorrect reading frame, the periodicity of 3 cannot be captured by DFA and the deviation value is small. For noncoding regions where the period-3 feature is usually lacking or weak, the PFD does not change much for the three reading frames. Note that the DFT magnitude for a coding region is similar for all three reading frames while the DFT phase is not. Codon indices based on the latter has improved performance compared with algorithms that use DFT magnitude (Kotlar and Lavner [20]). The PFD measure, whose value also varies for different reading frames, not only quantifies a sequence's coding strength well but also locates the reading frame correctly. The latter statement will be made more concrete shortly.

Algorithm for computing period-3 fractal deviations along a DNA sequence

Based on the observations above, we employ a sliding window technique (with window size w) to calculate PFD along a DNA sequence in a systematic fashion. The algorithm can be stated in four steps.

Step 1. Given a DNA sequence of length N, construct a DNA walk of length N based on the simple purine/pyrimidine rule (Peng et al [22]). Let w be the size of the sliding window. When successive windows overlap by $w - 1$ bases, a total of $N - w + 1$ windows can be

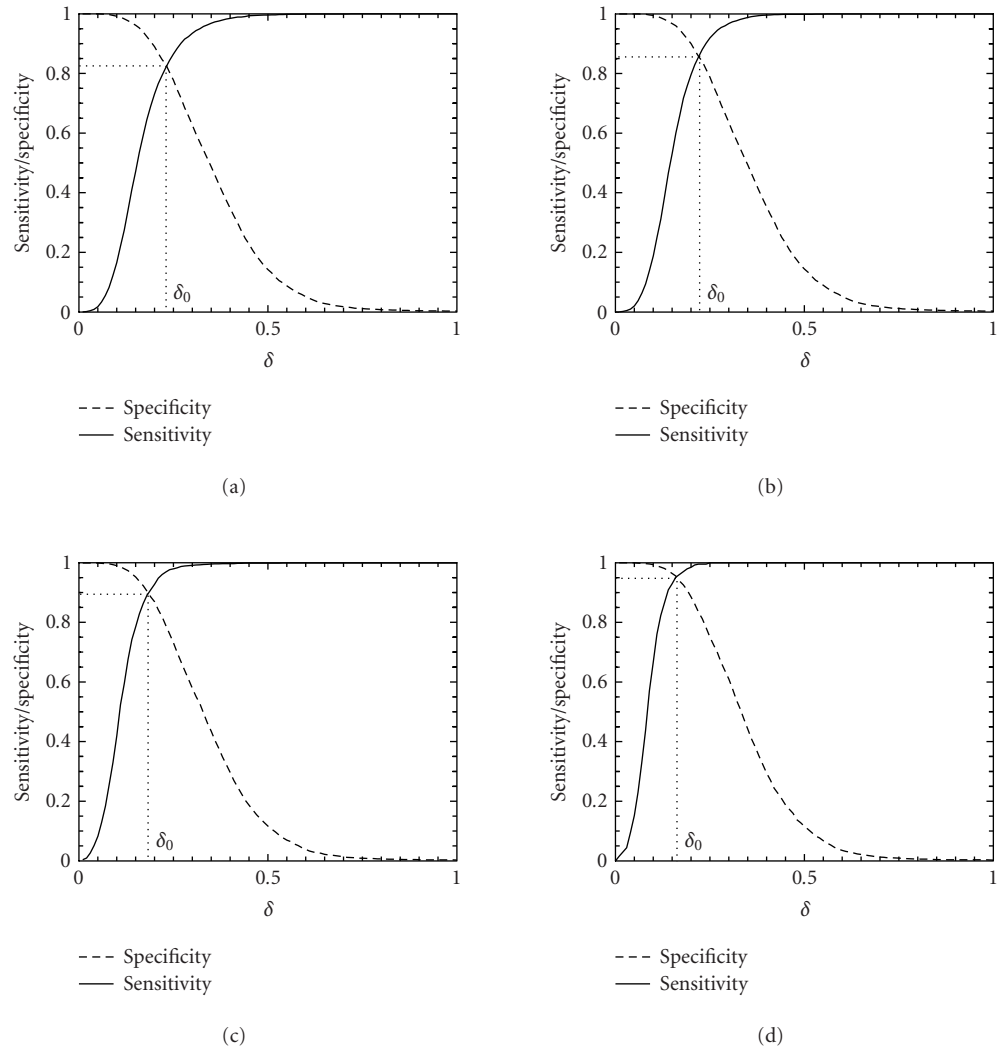


FIGURE 4. Distributions of MAXFD for the coding (solid curves) and noncoding (dashed curves) subsets of the 16 yeast chromosomes. The sliding window size is $w = 512$. The parameters n_1 and n_2 designate the coding/noncoding segments with lengths greater than n_1 and n_2, respectively. (a) $n_1 = n_2 = 1$, (b) $n_1 = n_2 = 256$, (c) $n_1 = n_2 = 512$, (d) $n_1 = n_2 = 1026$. See the text and Table 1 for more details.

obtained. For each window, the value of PFD can be computed based on (7).

Step 2. By common sense, one would associate each PFD for a window with the center of the window. In order to preserve information about the reading frames, however, this rule is slightly modified as follows: denote the position of the window along the DNA sequence by $[n, n + w - 1]$. We associate its PFD with the position $n + 3j$, where j is the largest integer such that $3j \leq w/2$.

Step 3. Form three reading-frame-specific deviation sequences by dividing the PFD(n) sequence into three subsets, $\mathrm{PFD}^1(1 + 3m)$, $\mathrm{PFD}^2(2 + 3m)$, $\mathrm{PFD}^3(3 + 3m)$, $m = 0, 1, 2, \ldots$, corresponding to the positions $(1, 4, 7, \ldots)$, $(2, 5, 8, \ldots)$, $(3, 6, 9, \ldots)$, respectively. For later convenience, we will denote $\mathrm{PFD}^1(1 + 3m)$, $\mathrm{PFD}^2(2 + 3m)$, $\mathrm{PFD}^3(3 + 3m)$ by $\mathrm{PFD}^1(m)$, $\mathrm{PFD}^2(m)$, $\mathrm{PFD}^3(m)$, $m = 0, 1, 2, \ldots$.

As we will illustrate in the next section, the above three steps automatically exhibit which reading frame is the correct one. Step 4 defines a simple but efficient codon index MAXFD.

Step 4. After PDF^i, $i = 1, 2, 3$ are obtained, we compute

$$\mathrm{MAXFD} = \frac{1}{\lfloor M/3 \rfloor} \sum_{m=1}^{\lfloor M/3 \rfloor} \max\left(\mathrm{PFD}^1(m), \mathrm{PFD}^2(m), \mathrm{PFD}^3(m)\right). \tag{8}$$

Let δ_0 be a threshold value. A segment under study is declared "coding" if the codon index MAXFD is greater than δ_0 and "noncoding" otherwise. In practice, the threshold

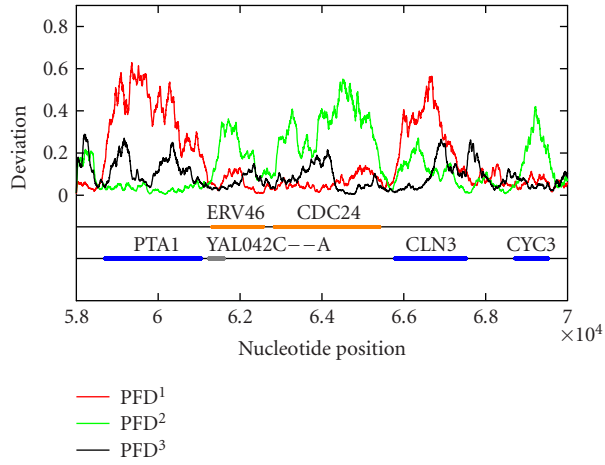


FIGURE 5. The reading-frame-specific PFD^i, $i = 1, 2, 3$ curves for a segment of DNA in yeast chromosome I (from nucleotide 58 000 to nucleotide 70 000). The sliding window size is $w = 512$. A 5th-order moving average filter has been applied. Colored horizontal bars on the two lines below the deviation curves are the open reading frames on the two strands of the chromosome, (first line: positive strand; second line: reverse strand). The orange and blue bars represent verified ORFs while a gray bar represents a dubious ORF.

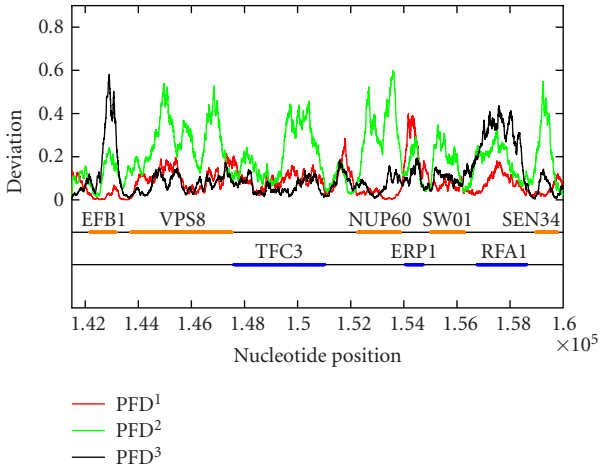


FIGURE 6. Period-3 fractal deviations (PFDs) of yeast chromosome I (a segment from nucleotide 141 500 to nucleotide 160 000).

value δ_0 is usually chosen to be where the cumulative distribution for the coding regions intersects with the complementary cumulative distribution for the noncoding regions (See Figure 4).

RESULTS AND DISCUSSION

The above algorithm has been used to calculate the PFD values of all of the 16 yeast chromosomes. To show that the algorithm is largely independent of the sliding window size as well as to show that the method is applicable to short DNA sequences, sliding window sizes of 128, 256, 512, and 1024 have been tried, with the best scaling regions identified as $[2^2, 2^4]$, $[2^2, 2^5]$, $[2^2, 2^6]$, and $[2^2, 2^7]$, respectively. For each window size, after a PFD sequence is obtained for an entire chromosome sequence, the three reading-frame-specific deviation sequences PFD$^i(m)$, $i = 1, 2, 3$, are plotted against their nucleotide positions along the chromosome in red, green, and black, respectively. For all four window sizes, the three deviation curves thus obtained exhibit similar and very interesting patterns. As examples, we have shown in Figures 5 and 6 the three reading-frame-specific PFD$^i(m)$ sequences for DNA segments in yeast chromosome I, from nucleotide 58 000 to nucleotide 70 000, and from 141 500 to nucleotide 160 000, respectively, where the window size w is chosen to be 512. To appreciate the correlations between the patterns of the variations of the PFD$^i(m)$ sequences and the coding/noncoding regions, the locations of the genes from both positive and reverse strands of the DNA sequence are also shown below the PFD$^i(m)$ curves. We observe a few interesting features. (i) Generally, the three curves, corresponding to three different colors, do not overlap with one another. This is a necessary condition for the three reading frames to be separable. (ii) In coding regions, both in the positive and the reverse strands, typically one of the three PFD^i curves displays a large value and separates considerably from the other two curves. By systematically comparing the yeast genome annotation data with the PFD^i curve with the largest values among the three, we have found that this is indeed the correct reading frame. Presumably, by searching for start and stop codons, one can aptly find out whether the gene is on the positive or the reverse strand of the genome, and determine which region(s) define(s) the gene. If this simple assumption would work, then a gene-finding algorithm that employs MAXFD as a codon index would require minimal amount of training. (iii) In noncoding regions, the three PFD^i curves are mixed. That means the three reading frames are more or less equivalent and inseparable.

We now evaluate the efficiency of the MAXFD as a codon index by studying all of the 16 yeast chromosomes.

Table 1. Accuracy of the PFD-based coding-region identification algorithm on different coding/noncoding subsets. The parameters N_1 and N_2 are the numbers of coding and noncoding sequences with length greater than n_1 and n_2, respectively. A DNA segment is declared "coding" if MAXFD $> \delta_0$ and "noncoding" otherwise. Accuracy is defined as the average of sensitivity and specificity. The threshold δ_0 is set at where sensitivity equals specificity.

n_1	N_1	n_2	N_2	δ_0	Sensitivity/specificity	
					$w = 512$	$w = 128$
1	4125	1	5993	0.1800	82.5%	84.7%
256	4067	256	4186	0.1660	85.7%	86.7%
512	3756	512	1948	0.1500	89.8%	89.4%
1026	2674	512	1948	0.1620	92.5%	91.2%
1026	2674	1026	650	0.1320	95.4%	94.4%

Our sample pool is comprised of two sets of DNA segments: the coding set, which contains 4125 verified ORFs (fully coding regions or exons), and the noncoding set, which contains 5993 segments (fully noncoding regions or introns). Different subsets of coding/noncoding segments are extracted according to the lengths of the sequence segments. These subsets are described by four parameters, N_1, n_1, N_2, n_2, where N_1 and N_2 are the numbers of coding and noncoding sequences with length greater than n_1 and n_2, respectively. After subsets of coding/noncoding sequences are chosen, MAXFD is then computed for all segments in those subsets. We denote the cumulative distribution of MAXFD over the two subsets as $P_C(\delta)$ and $P_{NC}(\delta)$. Then $1 - P_C(\delta)$ is the proportion of coding segments in C with MAXFD $> \delta$ and $P_{NC}(\delta)$ is the proportion of noncoding segments in NC with MAXFD $< \delta$. We define sensitivity as the proportion of segments in set C correctly labeled as "coding" and specificity as the proportion of segments in set NC correctly labeled as "noncoding." Given a threshold δ_0, the sensitivity and specificity are $1 - P_C(\delta_0)$ and $P_{NC}(\delta_0)$, respectively. If we define the percentage accuracy as the average of sensitivity and specificity, an optimal decision threshold is often set at where sensitivity equals specificity. By plotting $1 - P_C(f)$ and $P_{NC}(f)$ together, the optimal decision threshold is the abscissa of the point where the two curves intersect. The corresponding percentage accuracy is then simply $1 - P_C(\delta_0)$ (or $P_{NC}(\delta_0)$, since $P_{NC}(\delta_0) = 1 - P_C(\delta_0)$). Figure 4 shows the sensitivity/specificity curves for four configurations of (n_1, n_2). More detailed statistics for all five configurations of (n_1, n_2) studied are summarized in Table 1, for two window sizes, $w = 512$ and 128. With sliding window size $w = 512$, the percentage accuracy on the entire sample pool is 82.5%. When only those segments longer than the window size are concerned, the accuracy is increased to 89.8%. For coding and noncoding subsets with segment lengths greater than 1026, the accuracy is further improved to 95.4%. While one might think the statistics shown in Table 1 may become a lot worse when a much smaller sliding window size is used, this is not the case. In fact, when the sliding window size is reduced to 128, the accuracy for the long coding/noncoding sequences is only slightly degraded, while the accuracy for the entire coding/noncoding sequences is actually improved. Overall, we would conclude that the codon index proposed is fairly independent of the sliding window size.

In experiments involving expressed sequence tags (ESTs), the sequences available may all be short. Can the MAXFD index proposed still be useful? The answer is yes. When a sequence is very short, it is not necessary to use a sliding window to obtain three deviation curves. Instead one can simply obtain three values, PFD^1, PFD^2, PFD^3, from the sequence and find MAXFD using (8). If the value is very large, one has good reason to assume that the suspected EST indeed belongs to a coding region. Otherwise, it may not. When the former is the case, the reading frame with the largest PFD^i, where $i \in \{1, 2, 3\}$, very likely indicates the correct reading frame (assuming there is no error in the sequence). When the sequence under study is not too short, one can then employ the sliding window technique. The $\text{PFD}^1(m)$, $\text{PFD}^2(m)$, $\text{PFD}^3(m)$ curves that can be obtained this way will look like those obtained for a short segment of those shown in Figures 5 and 6. We note that the procedures outlined in this here have been applied to some experimentally obtained short DNA segments provided by Drs. Farmerie and Liu of the Institute of Biotechnology at the University of Florida.

ACKNOWLEDGMENT

J. B. Gao wishes to thank Drs. E. M. Marcotte, V. P. Roychowdhury, and I. Xenarios for many stimulating discussions.

REFERENCES

[1] Bennetzen JL, Hall BD. Codon selection in yeast. *J Biol Chem.* 1982;257(6):3026–3031.

[2] Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987;15(3):1281–1295.

[3] Jansen R, Bussemaker HJ, Gerstein M. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.* 2003;31(8):2242–2251.

[4] Zhang CT, Wang J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* 2000;28(14):2804–2814.
[5] Staden R, McLachlan AD. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 1982;10(1):141–156.
[6] Claverie JM, Bougueleret L. Heuristic informational analysis of sequences. *Nucleic Acids Res.* 1986;14(1):179–196.
[7] Farber R, Lapedes A, Sirotkin K. Determination of eukaryotic protein coding regions using neural networks and information theory. *J Mol Biol.* 1992;226(2):471–479.
[8] Fickett JW, Tung CS. Assessment of protein coding measures. *Nucleic Acids Res.* 1992;20(24):6441–6450.
[9] Fickett JW. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 1982;10(17):5303–5318.
[10] Shulman MJ, Steinberg CM, Westmoreland N. The coding function of nucleotide sequences can be discerned by statistical analysis. *J Theor Biol.* 1981;88(3):409–420.
[11] Borodovsky M, Koonin EV, Rudd KE. New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem Sci.* 1994;19(8):309–313.
[12] Almagor H. Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach. *J Theor Biol.* 1985;117(1):127–136.
[13] Silverman BD, Linsker R. A measure of DNA periodicity. *J Theor Biol.* 1986;118(3):295–300.
[14] Chechetkin VR, Turygin AY. Size-dependence of 3-periodicity and long-range correlations in DNA sequences. *Physics Lett A.* 1995;199(1-2):75–80.
[15] Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput Appl Biosci.* 1997;13(3):263–270.
[16] Trifonov EN. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A.* 1998;249:511–516.
[17] Yan M, Lin ZS, Zhang CT. A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics.* 1998;14:685–690.
[18] Anastassiou D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics.* 2000;16.(12):1073–1081.
[19] Issac B, Singh H, Kaur H, Raghava GP. Locating probable genes using Fourier transform approach. *Bioinformatics.* 2002;18(1):196–197.
[20] Kotlar D, Lavner Y. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 2003;13(8):1930–1937.
[21] Li W, Kaneko K. Long-range correlation and partial $1/f^a$ spectrum in a non-coding DNA sequence. *Europhys Lett.* 1992;17(7):655–660.
[22] Peng CK, Buldyrev SV, Goldberger AL, et al. Long-range correlations in nucleotide sequences. *Nature.* 1992;356(6365):168–170.
[23] Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys Rev Lett.* 1992;68(25):3805–3808.
[24] Mandelbrot BB. *The Fractal Geometry of Nature.* New York, NY: W. H. Freeman; 1982.
[25] Azbel MY. Random two-component one-dimensional Ising model for heteropolymer melting. *Phys Rev Lett.* 1973;31:589–592.
[26] Zhang CT, Zhang R, Ou HY. The Z curve database: a graphic representation of genome sequences. *Bioinformatics.* 2003;19(5):593–599.
[27] Berthelsen CL, Glazier JA, Skolnick MH. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys Rev A.* 1992;45(12):8902–8913.
[28] Cebrat S, Dudek MR, Gierlik A, Kowalczuk M, Mackiewicz P. Effect of replication on the third base of codons. *Physica A.* 1999;265(1-2):78–84.
[29] Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M. Scaling features of noncoding DNA. *Physica A.* 1999;273(1-2):1–18.
[30] Karlin S, Brendel V. Patchiness and correlations in DNA sequences. *Science.* 1993;259(5095):677–680.
[31] Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics.* 1994;49(2):1685–1689.
[32] Viswanathan GM, Buldyrev SV, Havlin S, Stanley HE. Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA. *Physica A.* 1998;249(1-4):581–586.
[33] Ossadnik SM, Buldyrev SV, Goldberger AL, et al. Correlation approach to identify coding regions in DNA sequences. *Biophys J.* 1994;67(1):64–70.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 147–154 • DOI: 10.1155/JBB.2005.147

RESEARCH ARTICLE

Classification and Selection of Biomarkers in Genomic Data Using LASSO

Debashis Ghosh[1] and Arul M. Chinnaiyan[2]

[1]*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA*
[2]*Departments of Pathology and Urology, University of Michigan, 1300 Catherine Road, Ann Arbor, MI 48109-1063, USA*

Received 3 June 2004; accepted 13 August 2004

High-throughput gene expression technologies such as microarrays have been utilized in a variety of scientific applications. Most of the work has been done on assessing univariate associations between gene expression profiles with clinical outcome (variable selection) or on developing classification procedures with gene expression data (supervised learning). We consider a hybrid variable selection/classification approach that is based on linear combinations of the gene expression profiles that maximize an accuracy measure summarized using the receiver operating characteristic curve. Under a specific probability model, this leads to the consideration of linear discriminant functions. We incorporate an automated variable selection approach using LASSO. An equivalence between LASSO estimation with support vector machines allows for model fitting using standard software. We apply the proposed method to simulated data as well as data from a recently published prostate cancer study.

INTRODUCTION

DNA microarrays simultaneously gauge the expression of thousands of genes in clinical samples. In this paper, we focus on cancer studies, where gene expression technologies have been applied extensively (Alizadeh et al [1]; Khan et al [2]; Dhanasekaran et al [3]). Obtaining large-scale gene expression profiles of tumors should theoretically allow for the identification of subsets of genes that function as prognostic disease markers or biologic predictors of therapeutic response. Because the data are highly multivariate and complex, it is important to develop automated statistical methods to detect systematic signals in gene expression patterns.

In cancer studies, analyses have typically focused on one of three problems. First, investigators have looked for genes that discriminate neoplastic from benign tissue. Statistically, this is the problem assessing differential expression of genes and has been studied by several authors; see, for example, Efron et al [4]. A second problem is clustering the samples to find subtypes of disease using algorithms such as those in [5]. The final class of problems is classification or supervised learning, which involves using the profile to predict some clinical outcome, such as the stage of disease. Suppose that in this instance, we treat the gene expression profile as the independent variables and tissue type as the response. A particular feature of microarray experiments is that the dimension of the predictor space (number of genes) is typically larger than the number of samples. This is known as the "large p, small n" paradigm (West [6]), so classification methods must take this into account.

One method to do this is apply prefiltering criteria in which the candidate number of genes for building a classifier is smaller than the number of samples. For example, Dudoit et al [7] performed a systematic comparison of several discrimination methods for the classification of tumors based on microarray experiments. However, they must perform an initial reduction in the number of predictors before building the classifier.

We wish to consider the joint effects of genes in determining classification rules for discriminating tumors. There are two assumptions that drive our proposed methodology. First, we assume that the joint effects of multiple genes must be considered in discriminating classes of disease. Recently, much attention has been given to the finding that a 70-gene signature can predict breast cancer survival (van't Veer et al [8]; van de Vijver et al [9]). However, most such gene signatures have been constructed using univariate methods. It seems reasonable to consider joint models, as genes are correlated because of their mutual involvement in disease pathways.

Correspondence and reprint requests to D. Ghosh, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA, E-mail: ghoshd@umich.edu

The second assumption is that there are individual genes that can discriminate classes. This is different from the latent factor and partial least squares proposals put forth by other authors (West [6]; Nguyen and Rocke [10]), where linear combinations of all available genes are used to predict the outcome. We seek to develop interpretable models for classification; for this purpose, using individual genes for predictors rather than linear combinations of genes seems reasonable.

In this paper, we develop classification rules based on the consideration of measures of diagnostic accuracy. In particular, we are interested in finding gene expression profiles that can discriminate between two populations. A unique challenge is posed because of the large p, small n problem. Our solution is to combine the problems of variable selection and classification. We suggest an approach for classification using the LASSO approach (Tibshirani [11]). An advantage of this approach is that some of the effects of the variables in these models are estimated to be exactly zero. These will represent genes that have no discriminatory power between the two classes, while those with nonzero coefficients will represent genes that can separate classes of tumors successfully. Thus, a by-product of the approach is the generation of a gene list. We exploit an equivalence between LASSO and support vector machines (SVMs) in order to fit the proposed classifier. The structure of the paper is as follows. In "materials and methods," we provide background on the data structures observed and the motivation based on biomarker combinations, which leads to the use of linear discriminant functions. We also provide a review of LASSO estimation (Tibshirani [11]) in this section. The latter two techniques are then involved in the proposed estimation procedure, described in "results and discussion." There, we also describe how to implement the proposed method using software for SVMs. Issues of model selection are also discussed. We describe the application of the proposed methodologies to simulated data and data from a recent cancer profiling study (Dhanasekaran et al [3]) in "prostate cancer gene expression data." Finally, some concluding remarks are made in "conclusion."

MATERIALS AND METHODS

Let $\mathbf{a}^T$ denote the transpose of the vector $\mathbf{a}$. For the ith sample ($i = 1, \ldots, n$), we let $\mathbf{X}_i = [X_{i1} \cdots X_{ip}]^T$ denote the $p \times 1$ gene expression profile vector (ie, X_{ij} is the gene expression measurement of the jth gene, $j = 1, \ldots, p$). We suppose that the data have already been preprocessed and normalized. In addition, it is assumed that the gene expression data are standardized so that for each gene, the mean is zero and standard deviation one. Let g_i denote the tumor class for the ith sample ($i = 1, \ldots, n$); we assume that there are two classes so that g_i takes values $g \in \{0, 1\}$. Here and in the sequel, we will refer to $g = 1$ as the diseased class and $g = 0$ as the healthy class; however, the methods proposed here are applicable to any two-class setting. In "LASSO estimation," we assume the existence of a continuous response variable Y_i for the ith sample ($i = 1, \ldots, n$).

ROC curves and optimal biomarker combinations

Our approach is to consider each measurement from a microarray for a single gene as a diagnostic test. Thus, for each subject, we have a high-dimensional vector of diagnostic test results. We then want to utilize this information in a way to separate the two populations of patients. This issue of finding combinations of biomarkers to accurately classify patients has been considered by Su and Liu [12], Baker [13], and Pepe and Thompson [14] in the statistical literature.

To combine information across the high-dimensional vector of gene expression profiles, we consider linear combinations of the form $\beta_0^T \mathbf{X}_i$, $i = 1, \ldots, n$. Without loss of generality, we will also assume that larger values of this linear combination corresponding to increasing likelihood of having $g = 1$. While the method can be easily extended to incorporate interactions between gene expression measurements, we focus on consideration of the main effects for purposes of exposition.

Suppose $\mathbf{X}^D$ represents the gene expression profile for a typical cancer specimen (ie, $g = 1$), and $\mathbf{X}^{\bar{D}}$ is the corresponding profile for a randomly chosen benign specimen. Note that in our situation, the diagnostic test is the linear combination $\beta_0^T \mathbf{X}$. One relevant quantity is the false positive rate based on a cutoff c, defined to be $FP(c) = P(\beta_0^T \mathbf{X} > c | g = 0)$. Similarly, the true positive rate is $TP(c) = P(\beta_0^T \mathbf{X} > c | g = 1)$. The true and false positive rates can be summarized by the receiver operating characteristic (ROC) curve, which is a graphical presentation of $\{FP(c), TP(c) : -\infty < c < \infty\}$. The ROC curve shows the tradeoff between increasing true and false positive rates. Tests that are have $\{FP(c), TP(c)\}$ values close to $(0, 1)$ indicate perfect discriminators, while those with $\{FP(c), TP(c)\}$ values close to the 45° line in the $(0, 1) \times (0, 1)$ plane are tests that are unable to discriminate between the diseased and healthy populations. Examples of ideal and noninformative ROC curves are given in Figures 1a and 1b.

While the specificity and sensitivity of a diagnostic test depend on the cutoff value chosen, a useful summary measure to consider is the area under the ROC curve. It can be shown mathematically that the area under curve is $P(\beta_0^T \mathbf{X}^D > \beta_0^T \mathbf{X}^{\bar{D}})$ (Bamber [15]). Under a binormal probability model, Su and Liu [12] showed that this quantity is optimized using the linear discriminant function. This motivates our choice of consideration of these variables. We next present an algorithm for estimation of these functions.

Linear discriminant functions by optimal scoring

While linear discriminant analysis (LDA) is typically calculated using matrix algebra techniques, an alternative method of calculating them is through the use of optimal scoring (Hastie et al [16, 17]). In this method, the

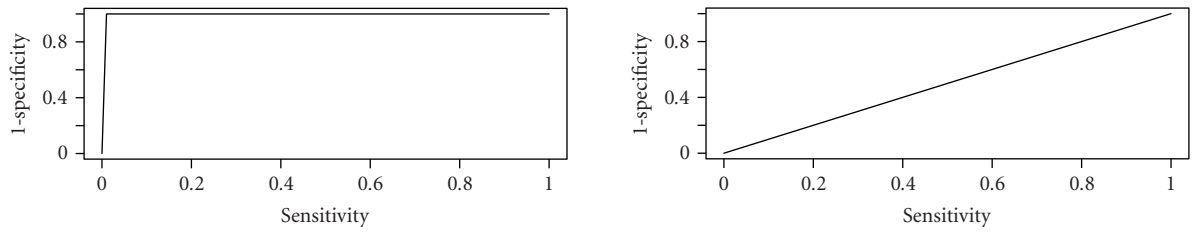


Figure 1. Receiver operating characteristic (ROC) curves for (a) ideal and (b) noninformative tests.

problem of classification into two groups is reexpressed as a regression problem based on quantities known as optimal scores.

The point of optimal scoring is to turn the categorical class labels into quantitative variables. Let $\boldsymbol{\theta}(g) = [\theta(g_1), \ldots, \theta(g_n)]^T$ be the $n \times 1$ vector of quantitative scores assigned to $\mathbf{g}$ for the kth class. The optimal scoring problem involves finding the vector of coefficients $\boldsymbol{\eta} \equiv (\eta_1, \eta_2, \ldots, \eta_p)$ and the scoring map $\theta : \{0, 1\} \to R$ that minimize the following average squared residual:

$$\text{ASR} = n^{-1} \sum_{i=1}^{n} \{\theta(g_i) - \mathbf{X}_i^T \boldsymbol{\eta}\}^2. \tag{1}$$

Let $\mathbf{Z}$ be an $n \times 2$ matrix with the ith row equal to $(1, 0)$ if $g_i = 1$ and $(0, 1)$ if $g_i = 0$ $(i = 1, \ldots, n)$. The optimal scores are assumed to be mutually orthogonal and normalized with respect to an inner product. Thus, the minimization of (1) is subject to the constraint $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\|^2 = 1$, where $\boldsymbol{\Theta} = [\theta(0) \;\; \theta(1)]^T$ is a 2×1 vector of the optimal scores. Hastie et al [16] state that the minimization of this constrained optimization problem leads to estimates of $\boldsymbol{\eta}$ that are proportional to the discriminant variables (ie, the discriminant function) in LDA. In particular, they propose the following algorithm for the estimation of the LDA functions

(1) Choose an initial score matrix $\boldsymbol{\Theta}_0$ satisfying $\boldsymbol{\Theta}_0^T \mathbf{D}_p \boldsymbol{\Theta}_0 = \mathbf{I}$, where $\mathbf{D}_p = \mathbf{Z}^T\mathbf{Z}/n$. Let $\boldsymbol{\Theta}_0^* = \mathbf{Z}\boldsymbol{\Theta}_0$.

(2) Let $\mathbf{X}$ be the $n \times p$ matrix with ith row $\mathbf{X}_i$. Fit a linear regression model of $\boldsymbol{\Theta}_0^*$ on $\mathbf{X}$, yielding fitted values $\widehat{\boldsymbol{\Theta}}$. Let $\widehat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.

(3) Obtain the eigenvector matrix $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}_0^{*T}\widehat{\boldsymbol{\Theta}}$; the optimal scores are then $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}_0\boldsymbol{\Phi}$.

(4) Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \boldsymbol{\Phi}^T\widehat{\mathbf{f}}(\mathbf{x})$.

As mentioned before, a problem with attempting to apply standard linear discriminant function methods to the data here is that there is not a numerically unique solution because p is larger than n. Thus, some type of regularization is needed. Our approach is based on the LASSO, which is described in the next section.

LASSO estimation

We suppose that our data are $(Y_i, \mathbf{X}_i)$, where Y_i $(i = 1, \ldots, n)$ is a continuous variable. The LASSO solution is to the optimization problem of minimizing

$$\sum_{i=1}^{n} (Y_i - \beta^T \mathbf{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ and $\lambda \geq 0$ is a penalty term. Thus, the constraint that is utilized is an L_1 constraint. An alternative way of formulating (2) is to minimize $\sum_{i=1}^{n}(Y_i - \beta^T\mathbf{X}_i)^2$, subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq t$. Note that in the absence of the constraint, the solution is given by the ordinary least squares (OLS) estimator. If the usual OLS estimator satisfies the constraint, then the LASSO and OLS estimates of β coincide. However, for smaller values of t, some of the components of β are estimated to be zero. In the linear regression setting, LASSO estimation has been considered by Tibshirani [11].

For a given value of t, minimization of $\sum_{i=1}^{n}(Y_i - \beta^T\mathbf{X}_i)^2$ subject to an L_1 constraint on the components of β is a quadratic programming problem with 2^p linear equality constraints. A sequential algorithm is given by Tibshirani [11] to solve the optimization problem.

While Tibshirani [11] considered estimating coefficients in regression models using LASSO, our interest is in using gene expression data to classify tumors. In particular, we seek to extend the LDA approach advocated by Dudoit et al [7] to handle the case where p is larger than n. We outline the proposed method in the next section.

Estimation methods

We propose to use an optimal scoring procedure for classification, where LASSO estimation is incorporated. In the notation of the previous section, we wish to solve the following optimization problem. Minimize

$$n^{-1} \sum_{i=1}^{n} \{\theta(g_i) - \mathbf{X}_i^T \boldsymbol{\eta}\}^2 + \lambda \sum_{j=1}^{p} |\eta_j| \tag{3}$$

subject to the constraint $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\|^2 = 1$. Here is the outline for our procedure.

(1) Choose an initial score matrix $\boldsymbol{\Theta}_0$ satisfying $\boldsymbol{\Theta}_0^T \mathbf{D}_p \boldsymbol{\Theta}_0 = \mathbf{I}$, and let $\boldsymbol{\Theta}_0 = \mathbf{Z}\boldsymbol{\Theta}$.

Table 1. Classification error rates (x 100) from simulation study. Numbers in parentheses represent standard errors associated with misclassification rates.

Sample size	$\pi = 0.05$ small effects	$\pi = 0.05$ large effects	$\pi = 0.5$ small effects	$\pi = 0.5$ large effects
$(n_0, n_1) = (15, 15)$	17.3 (1.65)	15.8 (1.63)	12.3 (1.21)	11.9 (1.30)
$(n_0, n_1) = (20, 10)$	20.7 (1.51)	19.3 (1.45)	13.3 (1.35)	12.7 (1.38)
$(n_0, n_1) = (50, 50)$	14.2 (1.15)	13.9 (1.24)	9.8 (1.02)	8.6 (1.11)
$(n_0, n_1) = (70, 30)$	18.3 (1.17)	17.6 (1.29)	10.2 (1.08)	9.9 (1.06)

(2) Fit a linear regression model of $\boldsymbol{\Theta}_0$ on $\mathbf{X}$ subject to an L_1 constraint on the parameters. Define the fitted values $\boldsymbol{\Theta}_0^*$. Let $\hat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.

(3) Obtain the eigenvector matrix $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}_0^{*T}\boldsymbol{\Theta}_0$; the optimal scores are $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0\boldsymbol{\Phi}$.

(4) Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \boldsymbol{\Phi}^T\hat{\mathbf{f}}(\mathbf{x})$.

Note that we are incorporating the LASSO estimation procedure in step (2) of the algorithm. We cannot use the algorithm of Tibshirani [11] because it is too computationally intensive for large p (number of genes). However, it turns out that the algorithm can be fit using standard software for SVMs, which we will now describe.

Support vector machines

An excellent descriptions of SVMs for classification can be found in [18]. We provide an overview of the method here. We assume that the data are $\{\mathbf{x}_i, y_i\}$ $(i = 1, \ldots, n)$, where $\mathbf{x}_i$ is a d-dimensional vector and $y_i \in \{-1, +1\}$ is the class label. The goal of SVMs is to find an optimal separating hyperplane between the observations with $y = -1$ and those with $y = 1$. This problem can be expressed as minimizing $\|\mathbf{w}\|^2$ subject to the following constraints:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq 1 - \xi_i \quad \text{for } y_i = 1, \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq 1 - \xi_i \quad \text{for } y_i = -1, \\ \xi_i &\geq 0 \quad \text{for } i = 1, \ldots, n. \end{aligned} \tag{4}$$

Details on how to solve the optimization problem can be found in [18, chapter 7]. In the unregularized case, fitting the LASSO model is equivalent to fitting an SVM classifier with the following $2p \times 1$ n-dimensional vectors as the inputs: $\mathbf{g}$, $\mathbf{Y}_k$ and $-\mathbf{Y}_k$ $(k = 1, \ldots, p)$, defined to be the sample labels, gene expression values and their negative values for the kth gene across the n samples. The label is the vector $\mathbf{y}_0$, defined to be -1 for the first entry and 1 for the other entries. The proof of the equivalence is given in the "appendix." We have created a macro in R (R foundation) that implements the proposed method and can be obtained from the first author.

As mentioned earlier, an advantage of this approach is that most of the gene effects are estimated to be exactly zero. The method can also identify the genes associated with each of the two classes. Genes whose coefficients are negative are associated with the class $g = -1$, while those with positive estimated coefficients are associated with $g = 1$.

As is evident in the algorithm from the previous section or in (3), the parameter λ needs to be estimated. We use fivefold cross-validation for this.

RESULTS AND DISCUSSION

Simulated data

We first performed a set of simulations to determine how well the proposed methods were at classification. We generated $p = 1000$ dimensional vectors for two populations. We considered the following sample size combinations $(n_0, n_1) = (15, 15)$, $(20, 10)$, $(50, 50)$, and $(70, 30)$, where n_k is the number of samples in the group with $g = k$ $(k = 0, 1)$. All the genes were assumed to be independent with a normal distribution and variance 1. We assumed a model in which a fraction π of the genes was differentially expressed between the two classes, $\pi = 0.05$ and $\pi = 0.5$ were considered. We examined two scenarios. For the first scenario, there was a big change in differential expression in the differentially expressed genes, a shift of 5 units in the mean. In the second scenario, the fold change was only a 1.5 unit difference in mean. For each simulation setting, 100 datasets were generated, and the classification error rates were estimated using three-fold cross-validation. No optimization was performed; we set $\lambda = 10$. The results are summarized in Table 1. Based on the table, we find that for larger sample sizes and larger effect sizes, as well as larger numbers of effects, the error rates are smaller.

However, in our simulations (data not shown), we found that the method had difficulty in selecting the correct variables when p is larger than n. This attests to the fact that variable selection in the situation of large p and small n is quite difficult. We discuss this situation in the "conclusion."

Table 2. List of genes underexpressed in prostate cancer relative to benign prostate tissue.

Clone ID	Gene name
Hs.288965	*Homo sapiens* cDNA: FLJ22300 fis, clone HRC04759
Hs.76307	Neuroblastoma, suppression of tumorigenicity 1
Hs.9615	Myosin, light polypeptide 9, regulatory
Hs.226795	Glutathione S-transferase pi
Hs.171731	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)

Prostate cancer gene expression data

The example we consider is from a prostate cancer study; a subset of the samples was considered by Dhanasekaran et al [3]. We focus here on noncancer versus cancer tissues. The samples are profiled using spotted cDNA (ie, red/green) microarrays; there are initially 101 samples profiled using 10 K chips (9984 genes). We have taken the following preprocessing steps:

(1) remove genes that are reported as missing in more than 10% of the samples;

(2) remove genes that have a variance less than 0.05 in all samples;

(3) impute measurements for missing genes using the median.

This leaves a total of 4880 genes for analysis.

We first performed an estimation of the error rate using fivefold cross-validation. This generally gave an error rate between 15–20% for various choices of λ, suggesting that the classifier is not sensitive to the choice of the smoothing parameter.

One of the by-products of the procedure is a list of genes that are estimated to have non-zero effects. We present the gene lists for $\lambda = 1$ in Table 2. Out of the 4880 genes, only 21 are estimated to have nonzero effects. Of the genes that are overexpressed in prostate cancer relative to benign prostate tissue, the early growth response (Hs. 326035/301865), feline sarcoma viral oncogene homolog (Hs.81665), T-cell receptor gamma locus (Hs. 112259), and fatty acid synthase (Hs.83190) have been seen by other investigators to be upregulated in prostate cancer, as in Table 3. The other genes on the list could represent false positives or genes whose joint effect is predictive of cancer status.

Conclusion

In this paper, we have introduced a new approach to the joint problems of classification and variable selection in the analysis of microarray data. These problems have been treated as separate problems in the previous literature. Our approach is combine the two problems by use of the LASSO.

This work has opened the way for several future avenues of research that we are currently investigating. First, a popular alternative to LDA in classification problems is logistic regression. It has been recently motivated by ROC considerations (McIntosh and Pepe [19]). While it is possible to formulate a LASSO estimation for logistic regression models, adapting the LASSO-SVM equivalence to this situation requires new algorithms. It will also be important to compare the performance of the two L_1-regularized procedures (LDA and logistic regression) on real and simulated microarray datasets.

In this paper, we focused on the two-class problem. While LDA and logistic regression can be extended to accommodate multicategorical responses, the ROC arguments that motivated the method here only exist for two populations. We are currently exploring theoretical frameworks for generalizing ROC ideas for multiple disease states.

The estimation procedure described in this paper allows the joint estimation of multivariate gene effects on the response (class label). The approach described here could be generalized by fitting more nonlinear gene effects in the estimation algorithm or by including higher-order interactions between genes. Another generalization is to perform a clustering of the genes and to enter the cluster averages as covariates in the model. Such an approach was taken by Hastie et al [20] and Tibshirani et al [21].

It is also of current interest to incorporate biological knowledge into microarray data analyses. In many instances, scientists are interested in the effects of a particular gene or pathway on genetic expression. In this context, approaches have been suggested by Zien et al [22] and Pavlidis et al [23] in which biological knowledge as represented by pathway scores or functional annotation status are correlated with gene expression. However, their approaches were univariate. There would be potential gains in efficiencies of analyses by considering joint models for pathways. We are currently studying the applicability of the joint estimation procedure described here to that setting.

Finally, a by-product of the method proposed here is that the individual genes can be estimated to have exactly zero effect on the response. The list of genes with estimated nonzero effects then comprise a gene list that

Table 3. List of genes overexpressed in prostate cancer relative to benign prostate tissue.

Clone ID	Gene name
Hs.326035/301865	Early growth response 1 -OR- dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
Hs.299221	Pyruvate dehydrogenase kinase, isoenzyme 4
Hs.81665	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
Hs.74267	Ribosomal protein L15
Hs.75431	Fibrinogen, gamma polypeptide
Hs.335797	ESTs, moderately similar to hypothetical protein FLJ20097 (*Homo sapiens*) (*H sapiens*)
Hs.82129	Carbonic anhydrase III, muscle specific
Hs.112259	T-cell receptor gamma locus
Hs.151258	Hypothetical protein FLJ21062
Hs.22394	Sec3-like
Hs.84190	Solute carrier family 19 (folate transporter), member 1
Hs.119597	Stearoyl-CoA desaturase (delta-9-desaturase)
Hs.131740	*Homo sapiens* cDNA FLJ30428 fis, clone BRACE2008941
Hs.50727	N-acetylglucosaminidase, alpha- (Sanfilippo disease IIIB)
Hs.83190	Fatty acid synthase
Hs.82961	*Homo sapiens*, clone MGC: 22588 IMAGE: 4696566, mRNA, complete cds

investigators can do further validation work on. However, in our simulations (data not shown), we found that the method had difficulty in selecting the correct v variables. This attests to the fact that variable selection in the situation of large p and small n is quite difficult. An alternative to the method proposed here is Bayesian variable selection methods (Lee et al [24]). We are currently exploring an adaptation of the algorithm described here to a Bayesian approach.

APPENDIX

If we let $\mathbf{w} = (w_1, \ldots, w_p)$, then SVMs can be shown to minimize $\|\mathbf{w}\|^2$ among all hyperplanes with norm 1, subject to the constraint that $g_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ for all $i = 1, \ldots, n$. The quantity $2/\|\mathbf{w}\|$ is known as the margin. In other words, we are trying to find the separating hyperplane that maximizes the margin among all classifiers that satisfy the inequality constraints. Using Lagrange multipliers, we can formulate the optimization problem as finding $\mathbf{w}$ and b to minimize

$$L(\mathbf{w}, b) \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \gamma_i g_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) + \boldsymbol{\gamma}' \mathbf{1}, \quad \text{(A.1)}$$

subject to $\gamma_i \geq 0$ $(i = 1, \ldots, n)$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$. Instead, we consider the dual of this problem, which is to maximize L such that the derivatives with respect to $\mathbf{w}$ and b vanish and also that $\gamma_i \geq 0$ $(i = 1, \ldots, n)$. By differentiating (A.1) with respect to $\mathbf{w}$ and b and setting the resulting derivatives equal to $\mathbf{0}$, we obtain

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} \gamma_i g_i \mathbf{x}_i = \mathbf{0}, \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^{n} \gamma_i g_i = 0. \end{aligned} \quad \text{(A.2)}$$

Equations (A.2) yield the solutions $\hat{\mathbf{w}} = \sum_{i=1}^{n} \gamma_i g_i \mathbf{x}_i$ and $\sum_{i=1}^{n} \gamma_i g_i = 0$. If we plug in the formula for $\hat{\mathbf{w}}$ into (A.1), the optimization problem becomes one of maximizing the dual function $W(\boldsymbol{\eta})$ over $\boldsymbol{\gamma} \geq \mathbf{0}$ and $\sum_{i=1}^{n} \gamma_i g_i = 0$, where

$$W(\boldsymbol{\eta}) = \sum_{j=1}^{n} \gamma_j - \frac{1}{2} \sum_{j,k=1}^{n} \gamma_j \gamma_k g_j g_k \langle \mathbf{x}_j, \mathbf{x}_k \rangle. \quad \text{(A.3)}$$

Tibshirani [11] considered the following estimation problem Minimize

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{Z}_i^T \beta\right)^2 \quad \text{(A.4)}$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$. Note that this minimization problem is equivalent to minimizing (A.4) subject to $\sum_{j=1}^{p} (\beta_j^+ + \beta_j^-) \leq t$, where $a^+ = \max(0, a)$ and $a^- = -\min(0, -a)$. We can equivalently consider minimization of

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} Z_{ij}\beta_j^+ + \sum_{j=1}^{p} Z_{ij}\beta_j^-\right)^2 - C\left[t - \sum_{j=1}^{p} \beta_j^+ - \sum_{j=1}^{p} \beta_j^-\right] \quad \text{(A.5)}$$

subject to $\beta_j^+ \geq 0$ and $\beta_j^- \geq 0$, $j = 1, \ldots, p$. We introduce some more notation. For $k = 1, \ldots, 2p$, define W_{ik} as Z_{ik} for $k = 1, \ldots, p$ and $-Z_{i(k-p-1)}$ for $k = p+1, \ldots, 2p$. Similarly, define the $2p \times 1$ dimensional vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{2p})$ by $\eta_j = \beta_j^+$ for $j = 1, \ldots, p$ and $\eta_j = \beta_{j-p-1}^-$ for $j = p+1, \ldots, 2p$. Thus, (A.5) can be written as

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{2p} W_{ij}\eta_j \right)^2 - C\left[t - \sum_{j=1}^{2p} \eta_j \right]. \tag{A.6}$$

The optimization problem now is to minimize (A.6) subject to $\eta_j \geq 0$ for $j = 1, \ldots, 2p$. Expanding the squared term in (A.6), we have

$$\sum_{i=1}^{n} \left(Y_i^2 - 2Y_i \sum_{j=1}^{2p} W_{ij}\eta_j - \sum_{j,k=1}^{2p} \eta_j\eta_k W_{ij}W_{ik} \right) - C\left[t - \sum_{j=1}^{2p} \eta_j \right]. \tag{A.7}$$

Distributing the summation sign and interchanging indices, (A.7) is equivalent to

$$\langle Y, Y \rangle - 2\sum_{j=1}^{2p} \langle W_j, Y \rangle \eta_j + \sum_{j,k=1}^{2p} \eta_j\eta_k \langle W_j, W_k \rangle - C\left[t - \sum_{j=1}^{2p} \eta_j \right]. \tag{A.8}$$

In particular, we want to minimize (A.8).

We now reconsider the optimization problem (A.3). Suppose we define new observations $(g_i, \mathbf{x}_i)$ $(i = 1, \ldots, 2p+1)$ by $g_1 = -1$ and $g_j = 1$ for $j = 2, \ldots, 2p+1$, $\mathbf{x}_1 = Y/t$, and $\mathbf{x}_j = W_{j-1}$ for $j = 2, \ldots, 2p+1$ and parameters $(\gamma_1, \ldots, \gamma_{2p+1})$ by

$$\gamma_1 = \frac{2t^2}{\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{2p} W_{ij}\eta_j \right)^2} \tag{A.9}$$

and $\gamma_j = \alpha_1\eta_{j-1}/t$ for $j = 2, \ldots, 2p+1$. Then the condition $\sum_{i=1}^{2p+1} \gamma_i g_i = 0$ is equivalent to $\gamma_1 = \sum_{i=2}^{2p+1} \gamma_i$, which after further algebraic simplification, yields $\sum_{j=1}^{2p} \eta_j = t$. Considerable algebraic simplification gives that maximizing (A.3) can be rewritten as a problem of maximizing

$$2\alpha_1 - \frac{1}{2}\frac{\alpha_1^2}{t^2}\langle Y, Y \rangle + \frac{\alpha_1^2}{t^2} \sum_{j=1}^{2p} \eta_j \langle W_j, Y \rangle - \frac{1}{2}\frac{\alpha_1^2}{t^2} \sum_{j,k=1}^{2p} \eta_j\eta_k g_j \langle W_j, W_k \rangle \tag{A.10}$$

subject to $\boldsymbol{\eta} \geq 0$ and $\sum_{j=1}^{2p} \eta_j = t$. Because $\alpha_1 \geq 0$, comparison of problems (A.10) and (A.8) reveal that they should yield the same solution.

ACKNOWLEDGMENT

The research of the first author was supported by grant NIH 1R01GM72007-01 from the Joint DMS/DBS/NIGMS Biological Mathematics Program.

REFERENCES

[1] Alizadeh AA, Ross DT, Perou CM, van de Rijn M. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol.* 2001;195(1):41–52.

[2] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.* 2001;7(6):673–679.

[3] Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature.* 2001;412(6849):822–826.

[4] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc.* 2001;96(456):1151–1160.

[5] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA.* 1998;95(25):14863–14868.

[6] West M. Bayesian factor regression models in the "large p, small n" paradigm. In: *Bayesian Statistics 7 Proceedings of the Seventh Valencia International Meeting.* New York, NY: Oxford University Press; 2003:723–732.

[7] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc.* 2002;97(457):77–87.

[8] van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530–536.

[9] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002;347(25):1999–2009.

[10] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics.* 2002;18(1):39–50.

[11] Tibshirani RJ. Regression shrinkage and selection via the LASSO. *J Roy Statist Soc B.* 1996;58(1):267–288.

[12] Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *J Am Stat Assoc.* 1993;88:1350–1355.

[13] Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics.* 2000;56(4):1082–1087.

[14] Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics.* 2000;1(2):123–140.

[15] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating

characteristic graph. *J Math Psych.* 1975;12(4):387–415.

[16] Hastie T, Tibshirani R, Buja A. Flexible discriminant analysis by optimal scoring. *J Am Stat Assoc.* 1994;89(428):1255–1270.

[17] Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. *Ann Statist.* 1995;23(1):73–102.

[18] Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge: Cambridge University Press; 2000.

[19] McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics.* 2002;58(3):657–664.

[20] Hastie T, Tibshirani R, Eisen MB, et al. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):Research0003. Epub 2000 Aug 04.

[21] Tibshirani R, Hastie T, Narasimhan B, et al. Exploratory screening of genes and clusters from microarray experiments. *Statist Sinica.* 2002;12(1):47–59.

[22] Zien A, Kuffner R, Zimmer R, Lengauer T. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:407–417.

[23] Pavlidis P, Lewis DP, Noble WS. Exploring gene expression data with class scores. *Pac Symp Biocomput.* 2002;7:474–485.

[24] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. *Bioinformatics.* 2003;19(1):90–97.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 155–159 • DOI: 10.1155/JBB.2005.155

RESEARCH ARTICLE

Gene Expression Data Classification With Kernel Principal Component Analysis

Zhenqiu Liu,[1] Dechang Chen,[2] and Halima Bensmail[3]

[1]*Bioinformatics Cell, US Army Medical Research and Materiel Command, 110 North Market Street, Frederick, MD 21703, USA*
[2]*Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA*
[3]*Department of Statistics, University of Tennessee, 331 Stokely Management Center, Knoxville, TN 37996, USA*

Received 3 June 2004; revised 28 August 2004; accepted 3 September 2004

One important feature of the gene expression data is that the number of genes M far exceeds the number of samples N. Standard statistical methods do not work well when $N < M$. Development of new methodologies or modification of existing methodologies is needed for the analysis of the microarray data. In this paper, we propose a novel analysis procedure for classifying the gene expression data. This procedure involves dimension reduction using kernel principal component analysis (KPCA) and classification with logistic regression (discrimination). KPCA is a generalization and nonlinear version of principal component analysis. The proposed algorithm was applied to five different gene expression datasets involving human tumor samples. Comparison with other popular classification methods such as support vector machines and neural networks shows that our algorithm is very promising in classifying gene expression data.

INTRODUCTION

One important application of gene expression data is the classification of samples into different categories, such as the types of tumor. Gene expression data are characterized by many variables on only a few observations. It has been observed that although there are thousands of genes for each observation, a few underlying gene components may account for much of the data variation. Principal component analysis (PCA) provides an efficient way to find these underlying gene components and reduce the input dimensions (Bicciato et al [1]). This linear transformation has been widely used in gene expression data analysis and compression (Bicciato et al [1], Yeung and Ruzzo [2]). If the data are concentrated in a linear subspace, PCA provides a way to compress data and simplify the representation without losing much information. However, if the data are concentrated in a nonlinear subspace, PCA will fail to work well. In this case, one may need to consider kernal principal component analysis (KPCA) (Rosipal and Trejo [3]). KPCA is a nonlinear version of PCA. It has been studied intensively in the last several years in the field of machine learning and has claimed success in many applications (Ng et al [4]). In this paper, we introduce a novel algorithm of classification, based on KPCA. Computational results show that our algorithm is effective in classifying gene expression data.

Correspondence and reprint requests to Zhenqiu Liu Bioinformatics Cell, U.S. Army Medical Research and Materiel Command, 110 North Market Street, Frederick, MD 21703, USA, E-mail: liu@stat.ohio-state.edu

ALGORITHM

A gene expression dataset with M genes (features) and N mRNA samples (observations) can be conveniently represented by the following gene expression matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix}, \quad (1)$$

where x_{li} is the measurement of the expression level of gene l in mRNA sample i. Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Mi})'$ denote the ith column (sample) of X with the prime $'$ representing the transpose operation, and y_i the corresponding class label (eg, tumor type or clinical outcome).

KPCA is a nonlinear version of PCA. To perform KPCA, one first transforms the input data $\mathbf{x}$ from the

original input space F_0 into a higher-dimensional feature space F_1 with the nonlinear transform $\mathbf{x} \to \Phi(\mathbf{x})$, where Φ is a nonlinear function. Then a kernel matrix K is formed using the inner products of new feature vectors. Finally, a PCA is performed on the centralized K, which is the estimate of the covariance matrix of the new feature vector in F_1. Such a linear PCA on K may be viewed as a nonlinear PCA on the original data. This property is sometimes called "kernel trick" in the literature. The concept of kernel is very important, here is a simple example to illustrate it. Suppose we have a two-dimensional input $\mathbf{x} = (x_1, x_2)'$, let the nonlinear transform be

$$\mathbf{x} \longrightarrow \Phi(\mathbf{x}) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)'. \tag{2}$$

Therefore, given two points $\mathbf{x}_i = (x_{i1}, x_{i2})'$ and $\mathbf{x}_j = (x_{j1}, x_{j2})'$, the inner product (kernel) is

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= \Phi(\mathbf{x}_i)'\Phi(\mathbf{x}_j) \\ &= x_{i1}^2x_{j1}^2 + x_{i2}^2x_{j2}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} \\ &\quad + 2x_{i1}x_{j1} + 2x_{i2}x_{j2} + 1 \\ &= \left(1 + x_{i1}x_{j1} + x_{i2}x_{j2}\right)^2 = \left(1 + \mathbf{x}_i'\mathbf{x}_j\right)^2, \end{aligned} \tag{3}$$

which is a second-order polynomial kernel. Equation (3) clearly shows that the kernel function is an inner product in the feature space and the inner products can be evaluated without even explicitly constructing the feature vector $\Phi(\mathbf{x})$.

The following are among the popular kernel functions:

(i) first norm exponential kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\beta||\mathbf{x}_i - \mathbf{x}_j||\right), \tag{4}$$

(ii) radial basis function (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{\sigma^2}\right), \tag{5}$$

(iii) power exponential kernel (a generalization of RBF kernel)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{r^2}\right)^{\beta}\right], \tag{6}$$

(iv) sigmoid kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\beta\mathbf{x}_i'\mathbf{x}_j\right), \tag{7}$$

(v) polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i'\mathbf{x}_j + p_2\right)^{p_1}, \tag{8}$$

where p_1 and $p_2 = 0, 1, 2, 3, \ldots$ are both integers.

For binary classification, our algorithm, based on KPCA, is stated as follows.

KPC classification algorithm

Given a training dataset $\{\mathbf{x}_i\}_{i=1}^n$ with class labels $\{y_i\}_{i=1}^n$ and a test dataset $\{\mathbf{x}_t\}_{t=1}^{n_t}$ with labels $\{y_t\}_{t=1}^{n_t}$, do the following.

(1) Compute the kernel matrix, for the training data, $K = [K_{ij}]_{n\times n}$, where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Compute the kernel matrix, for the test data, $K_{\text{te}} = [K_{ti}]_{n_t\times n}$, where $K_{ti} = K(\mathbf{x}_t, \mathbf{x}_i)$. K_{ti} projects the test data $\mathbf{x}_t$ onto training data $\mathbf{x}_i$ in the high-dimensional feature space in terms of the inner product.

(2) Centralize K using and K_{te}

$$\begin{aligned} K &= \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)K\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right), \\ K_{\text{te}} &= \left(K_{\text{te}} - \frac{1}{n}\mathbf{1}_{n_t}\mathbf{1}_n'K\right)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right). \end{aligned} \tag{9}$$

(3) Form an $n \times k$ matrix $Z = [z_1 \ z_2 \ \cdots \ z_k]$, where $z_1, z_2, \ldots, z_k$ are eigenvectors of K that correspond to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$. Also form a diagonal matrix D with λ_i in a position (i, i).

(4) Find the projections $\mathbf{V} = KZD^{-1/2}$ and $\mathbf{V}_{\text{te}} = K_{\text{te}}ZD^{-1/2}$ for the training and test data, respectively.

(5) Build a logistic regression model using $\mathbf{V}$ and $\{y_i\}_{i=1}^n$ and test the model performance using $\mathbf{V}_{\text{te}}$ and $\{y_t\}_{t=1}^{n_t}$.

We can show that the above KPC classification algorithm is a nonlinear version of the logistic regression. From our KPC classification algorithm, the probability of the label y, given the projection $\mathbf{v}$, is expressed as

$$P(y|\mathbf{w}, \mathbf{v}) = g\left(b + \sum_{i=1}^{k} w_iv_i\right), \tag{10}$$

where the coefficients $\mathbf{w}$ are adjustable parameters and g is the logistic function

$$g(u) = \left(1 + \exp(-u)\right)^{-1}. \tag{11}$$

Let n be the number of training samples and Φ the nonlinear transform function. We know each eigenvector z_i lies in the span of $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)$ for $i = 1, \ldots, n$ (Rosipal and Trejo [3]). Therefore one can write, for constants z_{ij},

$$z_i = z_{i1}\Phi(\mathbf{x}_1) + z_{i2}\Phi(\mathbf{x}_2) + \cdots + z_{in}\Phi(\mathbf{x}_n) = \sum_{j=1}^{n} z_{ij}\Phi(\mathbf{x}_j). \tag{12}$$

Given a test data $\mathbf{x}$, let v_i denote the projection of $\Phi(\mathbf{x})$ onto the ith nonlinear component with a normalizing factor $1/\sqrt{\lambda_i}$, we have

$$v_i = \frac{1}{\sqrt{\lambda_i}}\left(z_i'\Phi(\mathbf{x})\right) = \frac{1}{\sqrt{\lambda_i}}\sum_{j=1}^{n} z_{ij}K(\mathbf{x}_j, \mathbf{x}). \tag{13}$$

Substituting (13) into (10), we have

$$P(y|\mathbf{w},\mathbf{v}) = g\left(b + \sum_{j=1}^{n} c_j K(\mathbf{x}_j,\mathbf{x})\right), \tag{14}$$

where

$$c_j = \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} w_i z_{ij}, \quad i = 1,\ldots,n. \tag{15}$$

When $K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_j$, (14) becomes logistic regression. $K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_j$ is a linear kernel (polynomial kernel with $p_1 = 1$ and $p_2 = 0$). When we first normalize the input data through minusing their mean and then dividing their standard deviation, linear kernel matrix is the covariance matrix of the input data. Therefore KPC classification algorithm is a generalization of logistic regression.

Described in terms of binary classification, our classification algorithm can be readily employed for multiclass classification tasks. Typically, two-class problems tend to be much easier to learn than multiclass problems. While for two-class problems only one decision boundary must be inferred, the general c-class setting requires us to apply a strategy for coupling decision rules. For a c-class problem, we employ the standard approach where two-class classifiers are trained in order to separate each of the classes against all others. The decision rules are then coupled by voting, that is, sending the sample to the class with the largest probability.

Mathematically, we build c two-class classifiers based on a KPC classification algorithm in the form of (14) with the scheme "one against the rest":

$$p_i = P(y = i|\mathbf{x}) = g\left(b_i + \sum_{i=1}^{n} w_{ij} K(\mathbf{x}_i,\mathbf{x})\right), \tag{16}$$

where $i = 1, 2, \ldots, c$. Then for a test data point $\mathbf{x}_t$, we have the predicted class

$$\hat{y}_t = \arg\max_{i=1,\ldots,c} p_i(\mathbf{x}_t). \tag{17}$$

Feature and model selections

Since many genes show little variation across samples, gene (feature) selection is required. We chose the most informative genes with the highest likelihood ratio scores, described below (Ideker et al [5]). Given a two-class problem with an expression matrix $X = [x_{li}]_{M\times N}$, we have, for each gene l,

$$T(\mathbf{x}_l) = \log\frac{\sigma^2}{\sigma'^2}, \tag{18}$$

where

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{N} (x_{li} - \mu)^2, \\ \sigma'^2 &= \sum_{i\in \text{ class } 0} (x_{li} - \mu_0)^2 + \sum_{i\in \text{ class } 1} (x_{li} - \mu_1)^2. \end{aligned} \tag{19}$$

Here μ, μ_0, and μ_1 are the whole sample mean, the Class 0 mean, and the Class 1 mean, alternatively. We selected the most informative genes with the largest T values. This selection procedure is based on the likelihood ratio and used in our classification.

On the other hand, the dimension of projection (the number of eigenvectors) k used in the model can be selected based on Akaike's information criteria (AIC):

$$\text{AIC} = -2\log(\hat{L}) + 2(k+1), \tag{20}$$

where $\hat{L}$ is the maximum likelihood and k is the dimension of the projection in (10). The maximum likelihood $\hat{L}$ can also be calculated using (10):

$$\hat{L} = \prod_{i=1}^{n} (p(y|\mathbf{w},\mathbf{v}))^{y} (1 - p(y|\mathbf{w},\mathbf{v}))^{1-y}. \tag{21}$$

We can choose the best k with minimum AIC value.

COMPUTATIONAL RESULTS

To illustrate the applications of the algorithm proposed in the previous section, we considered five gene expression datasets: leukemia (Golub et al [6]), colon (Alon et al [7]), lung cancer (Garber et al [8]), lymphoma (Alizadeh et al [9]), and NCI (Ross et al [10]). The classification performance is assessed using the "leave-one-out (LOO) cross validation" for all of the datasets except for leukemia which uses one training and test data only. LOO cross validation provides more realistic assessment of classifiers which generalize well to unseen data. For presentation clarity, we give the number of errors with LOO in all of the figures and tables.

Leukemia

The leukemia dataset consists of expression profiles of 7129 genes from 38 training samples (27 ALL and 11 AML) and 34 testing samples (20 ALL and 14 AML). For classification of leukemia using a KPC classification algorithm, we chose the polynomial kernel $K(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^2$ and 15 eigenvectors corresponding to the first 15 largest eigenvalues with AIC. Using 150 informative genes, we obtained 0 training error and 1 test error. This is the best result compared with those reported in the literature. The plot for the output of the test data is given in Figure 1, which shows that all the test data points are classified correctly except for the last data point.

Colon

The colon dataset consists of expression profiles of 2000 genes from 22 normal tissues and 40 tumor samples. We calculated the classification result using a KPC classification algorithm with a kernel $K(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^2$. There were 150 selected genes and 25 eigenvectors selected with AIC criteria. The result is compared with that from the linear principal component (PC) logistic regression. The classification errors were calculated with the LOO

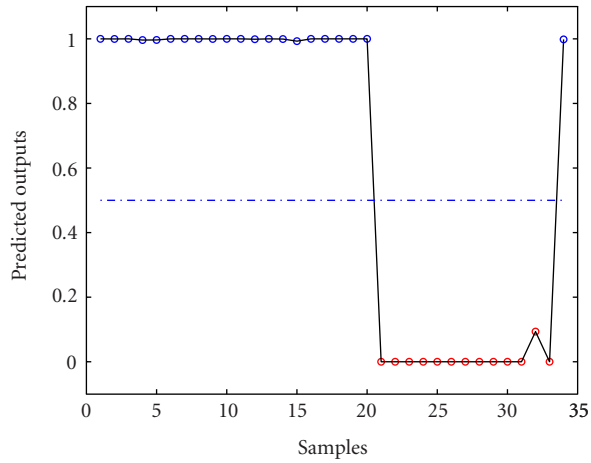


Figure 1. Output of the test data with KPC classification algorithm.

Table 1. Comparison for lung cancer.

Methods	Number of errors
KPC with a polynomial kernel	6
KPC with an RBF kernel	8
Linear PC classification	7
SVMs	7
Regularized logistic regression	12

method. The average error with linear PC logistic regression is 2 and the error with KPC classification is 0. The detailed results are given in Figure 2.

Lung cancer

The lung cancer dataset has 918 genes, 73 samples, and 7 classes. The number of samples per class for this dataset is small (less than 10) and unevenly distributed with 7 classes, which makes the classification task more challenging. A third-order polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^3$, and an RBF kernel with $\sigma = 1$ were used in the experiments. We chose the 100 most informative genes and 20 eigenvectors with our gene and model selection methods. The computational results of KPC classification and other methods are shown in Table 1. The results from SVMs for lung cancer, lymphoma, and NCI shown in this paper are those from Ding and Peng [11]. Six misclassifications with KPC and a polynomial kernel are given in Table 2. Table 1 shows that KPC with a polynomial kernel is performed better than that with an RBF kernel.

Lymphoma

The lymphoma dataset has 4026 genes, 96 samples, and 9 classes. A third-order polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^3$ and an RBF kernel with $\sigma = 1$ were used in our analysis. The 300 most informative genes and 21 eigenvectors corresponding to the largest eigenvalues were selected with the gene selection method and AIC criteria. A comparison of KPC with other methods is shown in Table 3.

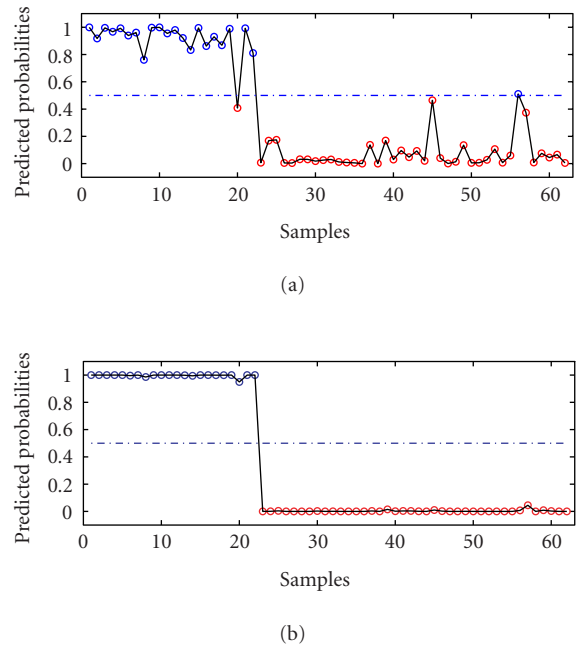


Figure 2. Outputs with (a) linear PC regression and (b) KPC classification.

Table 2. Misclassifications of lung cancer.

Sample index	True class	Predicted class
6	6	4
12	6	4
41	6	3
51	3	6
68	1	5
71	4	3

Table 3. Comparison for lymphoma.

Methods	Number of errors
KPC with a polynomial kernel	2
KPC with an RBF kernel	6
PC	5
SVMs	2
Regularized logistic regression	5

Misclassifications of lymphoma using KPC with a polynomial kernel are given in Table 4. There are only 2 misclassifications of class 1 using our KPC algorithm with a polynomial kernel, as shown in Table 4. The KPC with a polynomial kernel outperformed that with an RBF kernel in this experiment.

NCI

The NCI dataset has 9703 genes, 60 samples, and 9 classes. The third-order polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^3$ and an RBF kernel with $\sigma = 1$ were chosen in

Table 4. Misclassifications of lymphoma.

Sample index	True class	Predicted class
64	1	6
96	1	3

Table 5. Comparison for NCI.

Methods	Number of errors
KPC with a polynomial kernel	6
KPC with a RBF kernel	7
PC	6
SVMs	12
Logistic regression	6

Table 6. Misclassifications of NCI.

Sample index	True class	Predicted class
6	1	3
7	1	4
27	4	3
45	7	9
56	8	5
58	8	1

this experiment. The 300 most informative genes and 23 eigenvectors were selected with our simple gene selection method and AIC criteria. A comparison of computational results is summarized in Table 5 and the details of misclassification are listed in Table 6. KPC classification has equivalent performance with other popular tools.

DISCUSSIONS

We have introduced a nonlinear method, based on kPCA, for classifying gene expression data. The algorithm involves nonlinear transformation, dimension reduction, and logistic classification. We have illustrated the effectiveness of the algorithm in real life tumor classifications. Computational results show that the procedure is able to distinguish different classes with high accuracy. Our experiments also show that KPC classifications with second- and third-order polynomial kernels are usually performed better than that with an RBF kernel. This phenomena may be explained from the special structure of gene expression data. Our future work will focus on providing a rigorous theory for the algorithm and exploring the theoretical foundation that KPC with a polynomial kernel performed better than that with other kernels.

DISCLAIMER

The opinions expressed herein are those of the authors and do not necessarily represent those of the Uniformed Services University of the Health Sciences and the Department of Defense.

ACKNOWLEDGMENTS

D. Chen was supported by the National Science Foundation Grant CCR-0311252. The authors thank Dr. Hanchuan Peng, the Lawrence Berkeley National Laboratory for providing the NCI, lung cancer, and lymphoma data.

REFERENCES

[1] Bicciato S, Luchini A, Di Bello C. PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*. 2003;19(5):571–578.

[2] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics*. 2001;17(9):763–774.

[3] Rosipal R, Trejo LJ. Kernel partial least squares regression in RKHS: theory and empirical comparison. Tech. Rep. London: University of Paisley; March 2001.

[4] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems 14*, *Proceedings of the 2001*. Vancouver, British Columbia: MIT Press; 2001:849–856.

[5] Ideker T, Thorsson V, Siegel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol*. 2000;7(6):805–817.

[6] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(4539):531–537.

[7] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*. 1999;96(12):6745–6750.

[8] Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA*. 2001;98(24):13784–13789.

[9] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503–511.

[10] Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*. 2000;24(3):227–235.

[11] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. In: *Proc IEEE Bioinformatics Conference (CSB '03)*. Berkeley, Calif: IEEE; 2003:523–528.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 160–171 • DOI: 10.1155/JBB.2005.160

RESEARCH ARTICLE

Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree With Gene Selection

Yong Mao,[1] Xiaobo Zhou,[2] Daoying Pi,[1] Youxian Sun,[1] and Stephen T. C. Wong[2]

[1]*National Laboratory of Industrial Control Technology, Institute of Modern Control Engineering and College of Information Science and Engineering, Zhejiang University, Hangzhou 310027, China*
[2]*Harvard Center for Neurodegeneration & Repair and Brigham and Women's Hospital, Harvard Medical School, Harvard University, Boston, MA 02115, USA*

Received 3 June 2004; revised 2 November 2004; accepted 4 November 2004

We investigate the problems of multiclass cancer classification with gene selection from gene expression data. Two different constructed multiclass classifiers with gene selection are proposed, which are fuzzy support vector machine (FSVM) with gene selection and binary classification tree based on SVM with gene selection. Using F test and recursive feature elimination based on SVM as gene selection methods, binary classification tree based on SVM with F test, binary classification tree based on SVM with recursive feature elimination based on SVM, and FSVM with recursive feature elimination based on SVM are tested in our experiments. To accelerate computation, preselecting the strongest genes is also used. The proposed techniques are applied to analyze breast cancer data, small round blue-cell tumors, and acute leukemia data. Compared to existing multiclass cancer classifiers and binary classification tree based on SVM with F test or binary classification tree based on SVM with recursive feature elimination based on SVM mentioned in this paper, FSVM based on recursive feature elimination based on SVM can find most important genes that affect certain types of cancer with high recognition accuracy.

INTRODUCTION

By comparing gene expressions in normal and diseased cells, microarrays are used to identify diseased genes and targets for therapeutic drugs. However, the huge amount of data provided by cDNA microarray measurements must be explored in order to answer fundamental questions about gene functions and their interdependence [1], and hopefully to provide answers to questions like what is the type of the disease affecting the cells or which genes have strong influence on this disease. Questions like this lead to the study of gene classification problems.

Many factors may affect the results of the analysis. One of them is the huge number of genes included in the original dataset. Key issues that need to be addressed under such circumstances are the efficient selection of good predictive gene groups from datasets that are inherently noisy, and the development of new methodologies that can enhance the successful classification of these complex datasets.

For multiclass cancer classification and discovery, the performance of different discrimination methods including nearest-neighbor classifiers, linear discriminant analysis, classification trees, and bagging and boosting learning methods are compared in [2]. Moreover, this problem has been studied by using partial least squares [3], Bayesian probit regression [4], and iterative classification trees [5]. But multiclass cancer classification, combined with gene selection, has not been investigated intensively. In the process of multiclass classification with gene selection, where there is an operation of classification, there is an operation of gene selection, which is the focus in this paper.

In the past decade, a number of variable (or gene) selection methods used in two-class classification have been proposed, notably, the support vector machine (SVM) method [6], perceptron method [7], mutual-information-based selection method [8], Bayesian variable selection [2, 9, 10, 11, 12], minimum description

Correspondence and reprint requests to Stephen T. C. Wong, Harvard Center for Neurodegeneration & Repair and Brigham and Women's Hospital, Harvard Medical School, Harvard University, Boston, MA 02115, USA; stephen_wong@hms.harvard.edu

length principle for model selection [13], voting technique [14], and so on. In [6], gene selection using recursive feature elimination based on SVM (SVM-RFE) is proposed. When used in two-class circumstances, it is demonstrated experimentally that the genes selected by these techniques yield better classification performance and are biologically relevant to cancer than the other methods mentioned in [6], such as feature ranking with correlation coefficients or sensitivity analysis. But its application in multiclass gene selection has not been seen for its expensive calculation burden. Thus, gene preselection is adopted to get over this shortcoming; SVM-RFE is a key gene selection method used in our study.

As a two-class classification method, SVMs' remarkable robust performance with respect to sparse and noisy data makes them first choice in a number of applications. Its application in cancer diagnosis using gene profiles is referred to in [15, 16]. In the recent years, the binary SVM has been used as a component in many multiclass classification algorithms, such as binary classification tree and fuzzy SVM (FSVM). Certainly, these multiclass classification methods all have excellent performance, which benefit from their root in binary SVM and their own constructions. Accordingly, we propose two different constructed multiclass classifiers with gene selection: one is to use binary classification tree based on SVM (BCT-SVM) with gene selection while the other is FSVM with gene selection. In this paper, F test and SVM-RFE are used as our gene selection methods. Three groups of experiments are done, respectively, by using FSVM with SVM-RFE, BCT-SVM with SVM-RFE, and BCT-SVM with F test. Compared to the methods in [2, 3, 5], our proposed methods can find out which genes are the most important genes to affect certain types of cancer. In these experiments, with most of the strongest genes selected, the prediction error rate of our algorithms is extremely low, and FSVM with SVM-RFE shows the best performance of all.

The paper is organized as follows. Problem statement is given in "problem statement." BCT-SVM with gene selection is outlined in "binary classification tree based on SVM with gene" selection. FSVM with gene selection is described in "FSVM with gene selection." Experimental results on breast cancer data, small round blue-cell tumors data, and acute leukemia data are reported in "experimental results." Analysis and discussion are presented in "analysis and discussion." "Conclusion" concludes the paper.

PROBLEM STATEMENT

Assume there are K classes of cancers. Let $\mathbf{w} = [w_1, \ldots, w_m]$ denote the class labels of m samples, where $w_i = k$ indicates the sample i being cancer k, where $k = 1, \ldots, K$. Assume $x_1, \ldots, x_n$ are n genes. Let x_{ij} be the measurement of the expression level of the jth gene for the ith sample, where $j = 1, 2, \ldots, n$, $\mathbf{X} = [x_{ij}]_{m,n}$, denotes the expression levels of all genes, that is,

$$\mathbf{X} = \begin{bmatrix} \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene } n \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{12} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}. \tag{1}$$

In the two proposed methods, every sample is partitioned by a series of optimal hyperplanes. The optimal hyperplane means training data is maximally distant from the hyperplane itself, and the lowest classification error rate will be achieved when using this hyperplane to classify current training set. These hyperplanes can be modeled as

$$\boldsymbol{\omega}_{st}\mathbf{X}_i^T + b_{st} = 0 \tag{2}$$

and the classification functions are defined as $f_{st}(X_i^T) = \omega_{st}X_i^T + b_{st}$, where X_i denotes the ith row of matrix X; s and t mean two partitions which are separated by an optimal hyperplane, and what these partitions mean lies on the construction of multiclass classification algorithms; for example, if we use binary classification tree, s and t mean two halves separated in an internal node, which may be the root node or a common internal node; if we use FSVM, s and t mean two arbitrary classes in K classes. ω_{st} is an n-dimensional weight vector; b_{st} is a bias term.

SVM algorithm is used to determinate these optimal hyperplanes. SVM is a learning algorithm originally introduced by Vapnik [17, 18] and successively extended by many other researchers. SVMs can work in combination with the technique of "kernels" that automatically do a nonlinear mapping to a feature space so that SVM can settle the nonlinear separation problems. In SVM, a convex quadratic programming problem is solved and, finally, optimal solutions of ω_{st} and b_{st} are given. Detailed solution procedures are found in [17, 18].

Along with each binary classification using SVM, one operation of gene selection is done in advance. Specific gene selection methods used in our paper are described briefly in "experimental results." Here, gene selection is done before SVM trained means that when an SVM is trained or used for prediction, dimensionality reduction will be done on input data, X_i, referred to as the strongest genes selected. We use function $Y_i = I(\beta_{st}X_i^T)$ to represent this procedure, where β_{st} is an $n \times n$ matrix, in which only diagonal elements may be equal to 1 or 0; and all other elements are equal to 0; genes corresponding to the nonzero diagonal elements are important. β_{st} is gotten by specific gene selection methods; function $I(\cdot)$ means to select all nonzero elements in the input vector to construct a new vector , for example, $I([1\,0\,2])^T = [1\,2^T]$. So (2) is rewritten as

$$\boldsymbol{\beta}_{st}\mathbf{X}_i^T + b_{st} = 0, \qquad \mathbf{Y}_i = I\left(\boldsymbol{\beta}_{st}\mathbf{X}_i^T\right) \tag{3}$$

and the classification functions are rewritten as $f_{st}(X_i^T) = \beta_{st}X_i^T + b_{st}$ accordingly.

In order to accelerate calculation rate, preselecting genes before the training of multiclass classifiers is adopted. Based on all above, we propose two different constructed multiclass classifiers with gene selection: (1) binary classification tree based on SVM with gene selection, and (2) FSVM with gene selection.

BINARY CLASSIFICATION TREE BASED ON SVM WITH GENE SELECTION

Binary classification tree is an important class of machine-learning algorithms for multiclass classification. We construct binary classification tree with SVM; for short, we call it BCT-SVM. In BCT-SVM, there are $K-1$ internal nodes and K terminal nodes. When building the tree, the solution of (3) is searched by SVM at each internal node to separate the data in the current node into the left children node and right children node with appointed gene selection method, which is mentioned in "experimental results". Which class or classes should be partitioned into the left (or right) children node is decided at each internal node by impurity reduction [19], which is used to find the optimal construction of the classifier. The partition scheme with largest impurity reduction (IR) is optimal. Here, we use Gini index as our IR measurement criterion, which is also used in classification and regression trees (CARTs) [20] as a measurement of class diversity. Denote as M the training dataset at the current node, as M_L and M_R the training datasets at the left and right children nodes, as M_i sample set of class i in the training set, as $M_{R\cdot i}$ and $M_{L\cdot i}$ sample sets of class i of the training dataset at the left and right children nodes; and we use λ_Θ to denote the number of samples in dataset Θ; the current IR can be calculated as follows, in which c means the number of classes in the current node:

$$\mathrm{IR}(M) = \frac{1}{\lambda_M\lambda_{M_L}}\sum_{i=1}^{c}\left(\lambda_{M_{L\cdot i}}\right)^2 + \frac{1}{\lambda_M\lambda_{M_R}}\sum_{i=1}^{c}\left(\lambda_{M_{R\cdot i}}\right)^2 - \frac{1}{{\lambda_M}^2}\sum_{i=1}^{c}\left(\lambda_{M_i}\right)^2. \tag{4}$$

When the maximum of IR(M) is found out based on all potential combinations of classes in the current internal node, which part of data should be partitioned into the left children node is decided. For the details to construct the standard binary decision tree, we refer to [19, 20].

After this problem is solved, samples partitioned into the left children node are labeled with -1, and the others are labeled with 1, based on these measures, a binary SVM classifier with gene selection is trained using the data of the two current children nodes. As to gene selection, it is necessary because the cancer classification is a typical problem with small sample and large variables, and it will cause overfitting if we directly train the classifier with all genes; here, all gene selection methods based on two-class classfication could be used to construct β_{st} in (3). The process of building a whole tree is recursive, as seen in Figure 1.

When the training data at a node cannot be split any further, that node is identified as a terminal node and what we get from decision function corresponds to the label for a particular class. Once the tree is built, we could predict the results of the samples with genes selected by this tree; trained SVM will bring them to a terminal node, which has its own label. In the process of building BCT-SVM, there are $K-1$ operations of gene selection done. This is due to the construction of BCT-SVM, in which there are $K-1$ SVMs.

FSVM WITH GENE SELECTION

Other than BCT-SVM, FSVM has a pairwise construction, which means every hyperplane between two arbitrary classes should be searched using SVM with gene selection. These processes are modeled by (3).

FSVM is a new method firstly proposed by Abe and Inoue in [21, 22]. It was proposed to deal with unclassifiable regions when using one versus the rest or pairwise classification method based on binary SVM for $n(>2)$-class problems. FSVM is an improved pairwise classification method with SVM; a fuzzy membership function is introduced into the decision function based on pairwise classification. For the data in the classifiable regions, FSVM gives out the same classification results as pairwise classification with SVM method and for the data in the unclassifiable regions, FSVM generates better classification results than the pairwise classification with SVM method. In the process of being trained, FSVM is the same as the pairwise classification method with SVM that is referred to in [23].

In order to describe our proposed algorithm clearly, we denote four input variables: the sample matrix $\mathbf{X}_0 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_m\}^T$, that is, $\mathbf{X}_0$ is a matrix composed of some columns of original training dataset $\mathbf{X}$, which corresponds to preselected important genes; the class-label vector $\mathbf{y} = \{y_1, y_2, \ldots, y_k, \ldots, y_m\}^T$; the number of classes in training set ν; and the number of important genes used in gene selection κ. With these four input variables, the training process of FSVM with gene selection is expressed in (Algorithm 1).

In Algorithm 1, $\upsilon = GeneSelection(\mu, \phi, \kappa)$ is realization of a specific binary gene selection algorithm, υ denotes the genes important for two specific draw-out classes and is used to construct β_{st} in (3), $SVMTrain(\cdot)$ is realization of binary SVM algorithm, α is a Lagrange multiplier vector, and ϵ is a bias term. γ, $alpha$, and $bias$ are the output matrixes. γ is made up of all important genes selected, in which each row corresponds to a list of important genes selected between two specific classes. $alpha$ is a matrix with each row corresponding to Lagrange

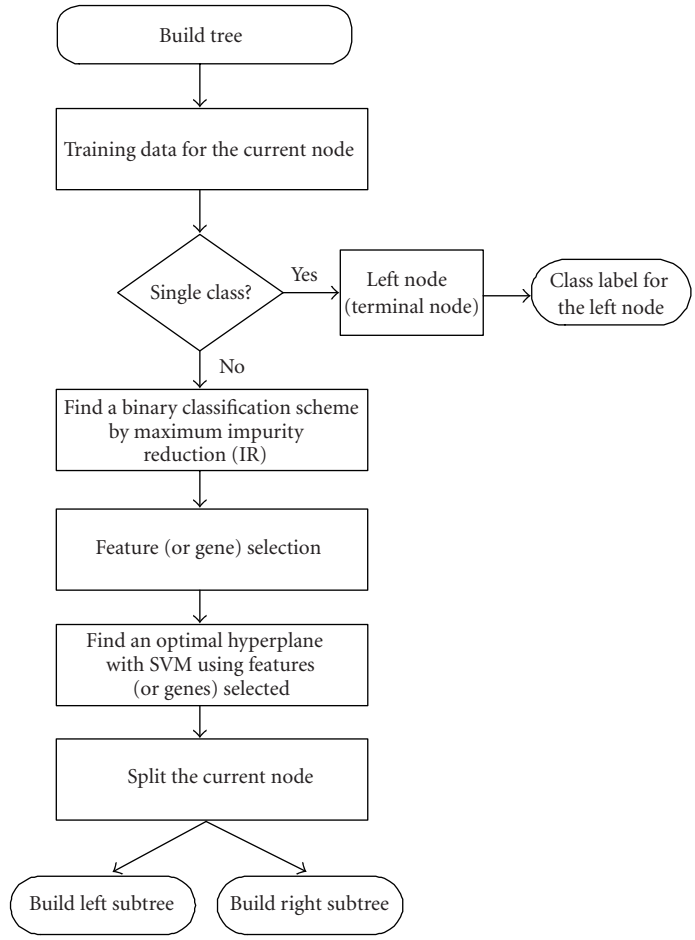


Figure 1. Binary classification tree based on SVM with gene selection.

multiplier vector by an SVM classifier trained between two specific classes, and *bias* is the vector made up of bias terms of these SVM classifiers.

In this process, we may see there are $K(K-1)/2$ SVMs trained and $K(K-1)/2$ gene selections executed. This means that many important genes relative to two specific classes of samples will be selected.

Based on the $K(K-1)/2$ optimal hyperplanes and the strongest genes selected, decision function is constructed based on (3). Define $f_{st}(X_i) = -f_{ts}(X_i), (s \neq t)$; the fuzzy membership function $m_{st}(X_i)$ is introduced on the directions orthogonal to $f_{st}(X_i) = 0$ as

$$m_{st}(\mathbf{X}_i) = \begin{cases} 1 & \text{for} \quad f_{st}(\mathbf{X}_i) \geq 1, \\ f_{st}(\mathbf{X}_i) & \text{otherwise.} \end{cases} \tag{5}$$

Using $m_{st}(X_i)(s \neq t, s = 1, \ldots, n)$, the class i membership function of X_i is defined as $m_s(X_i) = \min_{t=1,\ldots,n} m_{st}(X_i)$, which is equivalent to $m_s(X_i) = \min(1, \min_{s\neq t, t=1,\ldots,n} f_{st}(X_i))$; now an unknown sample X_i is classified by $arg\max_{s=1,\ldots,n} m_s(X_i)$.

EXPERIMENTAL RESULTS

F test and SVM-RFE are gene selection methods used in our experiments. In F test, the ratio $R(j) = \sum_{i=1}^{m}(\sum_{k=1}^{K} \mathbf{1}_{\Omega_i=k})(\overline{x}_{kj} - \overline{x}_j)^2 / \sum_{i=1}^{m}(\sum_{k=1}^{K} \mathbf{1}_{\Omega_i=k})(x_{ij} - \overline{x}_{kj})^2, 1 \leq j \leq n$, is used to select genes, in which $\overline{x}_j$ denotes the average expression level of gene j across all samples and $\overline{x}_{kj}$ denotes the average expression level of gene j across the samples belonging to class k where class k corresponds to $\{\Omega_i = k\}$; and the indicator function $\mathbf{1}_{\Omega}$ is equal to one if event Ω is true and zero otherwise. Genes with bigger $R(j)$ are selected. From the expression of $R(j)$, it can be seen F test could select genes among $l(>3)$ classes [14]. As to SVM-RFE, it is recursive feature elimination based on SVM. It is a circulation procedure for eliminating features combined with training an SVM classifier and, for each elimination operation, it consists of three steps: (1) train the SVM classifier, (2) compute the ranking criteria for all features, and (3) remove the feature with the smallest ranking scores, in which all ranking criteria are relative to the decision function of SVM. As a linear kernel SVM is used as a classifier

Inputs:
Sample matrix $\mathbf{X}_0 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_m\}^T$, class-label vector $\mathbf{y} = \{y_1, y_2, \ldots, y_k, \ldots, y_m\}^T$,
number of classes in training set $\nu = K$, and number of important genes we need $\kappa = z$

Initialize:
Set γ, *alpha*, and *bias* as empty matrixes. γ will be used to contain index number of ranked features;
alpha and *bias* will be used to contain parameters of FSVM

Training:
for $i \in \{1, \ldots, \nu - 1\}$
  for $j \in \{i - 1, \ldots, \nu\}$
    Initialize $\boldsymbol{\mu}$ as an empty matrix for containing draw-out samples and $\boldsymbol{\phi}$ as an empty vector for containing new-built class labels of class i and class j
    for $k \in \{1, \ldots, m\}$
      if $y_k = i$ *or* j
        Add $\mathbf{X}_0$' y_kth row to $\boldsymbol{\mu}$ as μ's last row
        if $y_k = i$, add element -1 to $\boldsymbol{\phi}$ as $\boldsymbol{\phi}$'s last element
        else, add element 1 to $\boldsymbol{\phi}$ as $\boldsymbol{\phi}$'s last element
      end
    end
    Gene selection
      Initialize $\boldsymbol{\upsilon}$ as an empty vector for containing important gene index number
      Get important genes between class i and class j
      $\boldsymbol{\upsilon} = GeneSelection(\boldsymbol{\mu}, \boldsymbol{\phi}, \kappa)$
      Put the results of gene selection into ranked feature matrix
      Add $\boldsymbol{\upsilon}$ to γ as γ's last row
    Train binary SVM using the row of genes selected right now
      Initialize $\boldsymbol{\tau}$ as an empty matrix for containing training data corresponding to the genes selected;
      Build the new matrix; Copy every column of $\boldsymbol{\mu}$ that $\boldsymbol{\upsilon}$ indicates into $\boldsymbol{\tau}$ as its column; Train the classifier
      $\{\alpha\ \epsilon\} = SVMTrain(\boldsymbol{\tau}, \boldsymbol{\phi})$
      Add α^T to *alpha* as *alpha*'s last row
      Add ϵ to *bias* as *bias*'s last element
  end
end

Outputs:
Ranked feature matrix γ
Two parameter matrixes of FSVM, *alpha* and *bias*

Algorithm 1. The FSVM with gene selection training algorithm.

between two specific classes s and t, the square of every element of weight vector ω_{st} in (2) is used as a score to evaluate the contribution of the corresponding genes. The genes with the smallest scores are eliminated. Details are referred to in [6]. To speed up the calculation, gene preselection is generally used. On every dataset we use the first important 200 genes are selected by F test before multiclass classifiers with gene selection are trained. Note that F test requires normality of the data to be efficient which is not always the case for gene expression data. That is the exact reason why we cannot only use F test to select genes. Since the P values of important genes are relatively low, that means the F test scores of important genes should be relatively high. Considering that the number of important genes is often among tens of genes, we preselect the number of genes as 200 according to our

experience in order to avoid losing some important genes. In the next experiments, we will show this procedure works effectively.

Combining these two specific gene selection methods with the multiclass classification methods, we propose three algorithms: (1) BCT-SVM with F test, (2) BCT-SVM with SVM-RFE, and (3) FSVM with SVM-RFE. As mentioned in [4, 9], every algorithm is tested with cross-validation (leave-one-out) method based on top 5, top 10, and top 20 genes selected by their own gene selection methods.

Breast cancer dataset

In our first experiment, we will focus on hereditary breast cancer data, which can be downloaded from the web page for the original paper [24]. In [24], cDNA microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. Twenty-two breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3226 genes for each tumor sample. We use our methods to classify BRCA1, BRCA2, and sporadic. The ratio data is truncated from below at 0.1 and above at 20.

Table 1 lists the top 20 strongest genes selected by using our methods. (For reading purpose, sometimes instead of clone ID, we use the gene index number in the database [24].) The clone ID and the gene description of a typical column of the top 20 genes selected by SVM-RFE are listed in Table 2; more information about all selected genes corresponding to the list in Table 1 could be found at http://www.sensornet.cn/fxia/top_20_genes.zip. It is seen that gene 1008 (keratin 8) is selected by all the three methods. This gene is also an important gene listed in [4, 7, 9]. Keratin 8 is a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry [24]. Gene 10 (phosphofructokinase, platelet) and gene 336 (transducer of ERBB2, 1) are also important genes listed in [7]. Gene 336 is selected by FSVM with SVM-RFE and BCT-SVM with SVM-RFE; gene 10 is selected by FSVM with SVM-RFE.

Using the top 5, 10, and 20 genes each for these three methods, the recognition accuracy is shown in Table 3. When using top 5 genes for classification, there is one error for BCT-SVM with F test and no error for the other two methods. When using top 10 and 20 genes, there is no error for all the three methods. Note that the performance of our methods is similar to that in [4], where the authors diagnosed the tumor types by using multinomial probit regression model with Bayesian gene selection. Using top 10 genes, they also got zero misclassification.

Small round blue-cell tumors

In this experiment, we consider the small round blue-cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing sarcoma (EWS) in [25]. The dataset of the four cancers is composed of 2308 genes and 63 samples, where the NB has 12 samples; the RMS has 23 samples; the NHL has 8 samples, and the EMS has 20 samples. We use our methods to classify the four cancers. The ratio data is truncated from below at 0.01.

Table 4 lists the top 20 strongest genes selected by using our methods. The clone ID and the gene description of a typical column of the top 20 genes selected by SVM-RFE are listed in Table 5; more information about all selected genes corresponding to the list in Table 4 could be found at http://www.sensornet.cn/fxia/top_20_genes.zip. It is seen that gene 244 (clone ID 377461), gene 2050 (clone ID 295985), and gene 1389 (clone ID 770394) are selected by all the three methods, and these genes are also important genes listed in [25]. Gene 255 (clone ID 325182), gene 107 (clone ID 365826), and gene 1 (clone ID 21652, (catenin alpha 1)) selected by BCT-SVM with SVM-RFE and FSVM with SVM-RFE are also listed in [25] as important genes.

Using the top 5, 10, and 20 genes for these three methods each, the recognition accuracy is shown in Table 6. When using top 5 genes for classification, there is one error for BCT-SVM with F test and no error for the other two methods. When using top 10 and 20 genes, there is no error for all the three methods.

In [26], Yeo et al applied k nearest neighbor (kNN), weighted voting, and linear SVM in one-versus-rest fashion to this four-class problem and compared the performances of these methods when they are combined with several feature selection methods for each binary classification problem. Using top 5 genes, top 10 genes, or top 20 genes, kNN, weighted voting, or SVM combined with all the three feature selection methods, respectively, without rejection all have errors greater than or equal to 2. In [27], Lee et al used multicategory SVM with gene selection. Using top 20 genes, their recognition accuracy is also zero misclassification number.

Acute leukemia data

We have also applied the proposed methods to the leukemia data of [14], which is available at http://www.sensornet.cn/fxia/top_20_genes.zip. The microarray data contains 7129 human genes, sampled from 72 cases of cancer, of which 38 are of type B cell ALL, 9 are of type T cell ALL, and 25 of type AML. The data is preprocessed as recommended in [2]: gene values are truncated from below at 100 and from above at 16 000; genes having the ratio of the maximum over the minimum less than 5 or the difference between the maximum and the minimum less than 500 are excluded; and finally the base-10 logarithm is applied to the 3571 remaining genes. Here we study the 38 samples in training set, which is composed of 19 B-cell ALL, 8 T-cell ALL, and 11 AML.

Table 1. The index no of the strongest genes selected in hereditary breast cancer dataset.

No	FSVM with SVM-RFE			BCT-SVM with F test		BCT-SVM with SVM-RFE	
	1	2	3	1	2	1	2
1	1008	1859	422	501	1148	750	1999
2	955	1008	2886	2984	838	860	3009
3	1479	10	343	3104	1859	1008	158
4	2870	336	501	422	272	422	2761
5	538	158	92	2977	1008	2804	247
6	336	1999	3004	2578	1179	1836	1859
7	3154	247	1709	3010	1065	3004	1148
8	2259	1446	750	2804	2423	420	838
9	739	739	2299	335	1999	1709	1628
10	2893	1200	341	2456	2699	3065	1068
11	816	2886	1836	1116	1277	2977	819
12	2804	2761	219	268	1068	585	1797
13	1503	1658	156	750	963	1475	336
14	585	560	2867	2294	158	3217	2893
15	1620	838	3104	156	609	501	2219
16	1815	2300	1412	2299	1417	146	585
17	3065	538	3217	2715	1190	343	1008
18	3155	498	2977	2753	2219	1417	2886
19	1288	809	1612	2979	560	2299	36
20	2342	1092	2804	2428	247	2294	1446

Table 2. A part of the strongest genes selected in hereditary breast cancer dataset (the first row of genes in Table 1).

Rank	Index no	Clone ID	Gene description
1	1008	897781	Keratin 8
2	955	950682	Phosphofructokinase, platelet
3	1479	841641	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)
4	2870	82991	Phosphodiesterase I/nucleotide pyrophosphatase 1 (homologous to mouse Ly-41 antigen)
5	538	563598	Human GABA-A receptor π subunit mRNA, complete cds
6	336	823940	Transducer of ERBB2, 1
7	3154	135118	GATA-binding protein 3
8	2259	814270	Polymyositis/scleroderma autoantigen 1 (75kd)
9	739	214068	GATA-binding protein 3
10	2893	32790	mutS (*E coli*) homolog 2 (colon cancer, nonpolyposis type 1)
11	816	123926	Cathepsin K (pycnodysostosis)
12	2804	51209	Protein phosphatase 1, catalytic subunit, beta isoform
13	1503	838568	Cytochrome c oxidase subunit VIc
14	585	293104	Phytanoyl-CoA hydroxylase (Refsum disease)
15	1620	137638	ESTs
16	1815	141959	*Homo sapiens* mRNA; cDNA DKFZp566J2446 (from clone DKFZp566J2446)
17	3065	199381	ESTs
18	3155	136769	TATA box binding protein (TBP)-associated factor, RNA polymerase II, A, 250kd
19	1288	564803	Forkhead (drosophila)-like 16
20	2342	284592	Platelet-derived growth factor receptor, alpha polypeptide

Table 3. Classifiers' performance on hereditary breast cancer dataset by cross-validation (number of wrong classified samples in leave-one-out test).

Classification method	Top 5	Top 10	Top 20
FSVM with SVM-RFE	0	0	0
BCT-SVM with F test	1	0	0
BCT-SVM with SVM-RFE	0	0	0

Table 4. The index no of the strongest genes selected in small round blue-cell tumors dataset.

No	FSVM with SVM-RFE						BCT-SVM with F test			BCT-SVM with SVM-RFE		
	1	2	3	4	5	6	1	2	3	1	2	3
1	246	255	1954	851	187	1601	1074	169	422	545	174	851
2	1389	867	1708	846	509	842	246	1055	1099	1389	1353	846
3	851	246	1955	1915	2162	1955	1708	338	758	2050	842	1915
4	1750	1389	509	1601	107	255	1389	422	1387	1319	1884	1601
5	107	842	2050	742	758	2046	1954	1738	761	1613	1003	742
6	2198	2050	545	1916	2046	1764	607	1353	123	1003	707	1916
7	2050	365	1389	2144	2198	509	1613	800	84	246	1955	2144
8	2162	742	2046	2198	2022	603	1645	714	1888	867	2046	2198
9	607	107	348	1427	1606	707	1319	758	951	1954	255	1427
10	1980	976	129	1	169	174	566	910	1606	1645	169	1
11	567	1319	566	1066	1	1353	368	2047	1914	1110	819	1066
12	2022	1991	246	867	1915	169	1327	2162	1634	368	509	867
13	1626	819	1207	788	788	1003	244	2227	867	129	166	788
14	1916	251	1003	153	1886	742	545	2049	783	348	1207	153
15	544	236	368	1980	554	2203	1888	1884	2168	365	603	1980
16	1645	1954	1105	2199	1353	107	2050	1955	1601	107	796	2199
17	1427	1708	1158	783	338	719	430	1207	335	1708	1764	783
18	1708	1084	1645	1434	846	166	365	326	1084	187	719	1434
19	2303	566	1319	799	1884	1884	1772	796	836	1626	107	799
20	256	1110	1799	1886	2235	1980	1298	230	849	1772	2203	1886

Table 7 lists the top 20 strongest genes selected by using our methods. The clone ID and the gene description of a typical column of the top 20 genes selected by SVM-RFE are listed in Table 8; more information about all selected genes corresponding to the list in Table 7 could be found at http://www.sensornet.cn/fxia/top_20_genes.zip. It is seen that gene 1882 (CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)), gene 4847 (zyxin), and gene 4342 (TCF7 transcription factor 7 (T cell specific)) are selected by all the three methods. In the three genes, the first two are the most important genes listed in many literatures. Gene 2288 (DF D component of complement (adipsin)) is another important gene having biological significance, which is selected by FSVM with SVM-RFE.

Using the top 5, 10, and 20 genes for these three methods each, the recognition accuracy is shown in Table 9. When using top 5 genes for classification, there is one error for FSVM with SVM-RFE, two errors for BCT-SVM with SVM-RFE and BCT-SVM with F test, respectively. When using top 10 genes for classification, there is no error for FSVM with SVM-RFE, two errors for BCT-SVM with SVM-RFE and four errors for BCT-SVM with F test. When using top 20 genes for classification, there is one error for FSVM with SVM-RFE, two errors for BCT-SVM with SVM-RFE and two errors for BCT-SVM with F test. Again note that the performance of our methods is similar to that in [4], where the authors diagnosed the tumor types by using multinomial probit regression model with Bayesian gene selection. Using top 10 genes, they also got zero misclassification.

ANALYSIS AND DISCUSSION

According to Tables 1–9, there are many important genes selected by these three multiclass classification algorithms with gene selection. Based on these selected genes, the prediction error rate of these three algorithms is low.

Table 5. A part of the strongest genes selected in small round blue-cell tumors dataset (the first row of genes in Table 4).

Rank	Index no	Clone ID	Gene description
1	246	377461	Caveolin 1, caveolae protein, 22kd
2	1389	770394	Fc fragment of IgG, receptor, transporter, alpha
3	851	563673	Antiquitin 1
4	1750	233721	Insulin-like growth factor binding protein 2 (36kd)
5	107	365826	Growth arrest-specific 1
6	2198	212542	*H sapiens* mRNA; cDNA DKFZp586J2118 (from clone DKFZp586J2118)
7	2050	295985	ESTs
8	2162	308163	ESTs
9	607	811108	Thyroid hormone receptor interactor 6
10	1980	841641	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)
11	567	768370	tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)
12	2022	204545	ESTs
13	1626	811000	Lectin, galactoside-binding, soluble, 3 binding protein (galectin 6 binding protein)
14	1916	80109	Major histocompatibility complex, class II, DQ alpha 1
15	544	1416782	Creatine kinase, brain
16	1645	52076	Olfactomedinrelated ER localized protein
17	1427	504791	Glutathione S-transferase A4
18	1708	43733	Glycogenin 2
19	2303	782503	*H sapiens* clone 23716 mRNA sequence
20	256	154472	Fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)

Table 6. Classifiers' performance on small round blue-cell tumors dataset by cross-validation (number of wrong classified samples in leave-one-out test).

Classification method	Top 5	Top 10	Top 20
FSVM with SVM-RFE	0	0	0
BCT-SVM with F test	1	0	0
BCT-SVM with SVM-RFE	0	0	0

By comparing the results of these three algorithms, we consider that FSVM with SVM-RFE algorithm generates the best results. BCT-SVM with SVM-RFE and BCT-SVM with F test have the same multiclass classification structure. The results of BCT-SVM with SVM-RFE are better than those of BCT-SVM with F test, because their gene selection methods are different; a better gene selection method combined with the same multiclass classification method will perform better. It means SVM-RFE is better than F test combined with multiclass classification methods; the results are similar to what is mentioned in [6], in which the two gene selection methods are combined with two-class classification methods.

FSVM with SVM-RFE and BCT-SVM with SVM-RFE have the same gene selection methods. The results of FSVM with SVM-RFE are better than those of BCT-SVM with SVM-RFE whether in gene selection or in recognition accuracy, because the constructions of their multiclass classification methods are different, which is explained in two aspects. (1) The genes selected by FSVM with SVM-RFE are more than those of BCT-SVM with SVM-REF. In FSVM there are $K(K-1)/2$ operations of gene selection, but in BCT-SVM there are only $K-1$ operations of gene selection. An operation of gene selection between every two classes is done in FSVM with SVM-RFE; (2) FSVM is an improved pairwise classification method, in which the unclassifiable regions being in BCT-SVM are classified by FSVM's fuzzy membership function [21, 22]. So, FSVM with SVM-RFE is considered as the best of the three.

CONCLUSION

In this paper, we have studied the problem of multiclass cancer classification with gene selection from gene expression data. We proposed two different new constructed classifiers with gene selection, which are FSVM with gene selection and BCT-SVM with gene

Table 7. The index no of the strongest genes selected in acute leukemia dataset.

No	FSVM with SVM-RFE			BCT-SVM with F test		BCT-SVM with SVM-RFE	
	1	2	3	1	2	1	2
1	6696	1882	6606	2335	4342	1882	4342
2	6606	4680	6696	4680	4050	6696	4050
3	4342	6201	4680	2642	1207	5552	5808
4	1694	2288	4342	1882	6510	6378	1106
5	1046	6200	6789	6225	4052	3847	3969
6	1779	760	4318	4318	4055	5300	1046
7	6200	2335	1893	5300	1106	2642	6606
8	6180	758	1694	5554	1268	2402	6696
9	6510	2642	4379	5688	4847	3332	2833
10	1893	2402	2215	758	5543	1685	1268
11	4050	6218	3332	4913	1046	4177	4847
12	4379	6376	3969	4082	2833	6606	6510
13	1268	6308	6510	6573	4357	3969	2215
14	4375	1779	2335	6974	4375	6308	1834
15	4847	6185	6168	6497	6041	760	4535
16	6789	4082	2010	1078	6236	2335	1817
17	2288	6378	1106	2995	6696	2010	4375
18	1106	4847	5300	5442	1630	6573	5039
19	2833	5300	4082	2215	6180	4586	4379
20	6539	1685	1046	4177	4107	2215	5300

Table 8. A part of the strongest genes selected in small round blue-cell tumors dataset (the second row of genes in Table 4).

Rank	Index no	Gene accession number	Gene description
1	1882	M27891_at	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)
2	4680	X82240_rna1_at	TCL1 gene (T-cell leukemia) extracted from *H sapiens* mRNA for T-cell leukemia/lymphoma 1
3	6201	Y00787_s_at	Interleukin-8 precursor
4	2288	M84526_at	DF D component of complement (adipsin)
5	6200	M28130_rna1_s_at	Interleukin-8 (IL-8) gene
6	760	D88422_at	Cystatin A
7	2335	M89957_at	IGB immunoglobulin-associated beta (B29)
8	758	D88270_at	GB DEF = (lambda) DNA for immunoglobin light chain
9	2642	U05259_rna1_at	MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)
10	2402	M96326_rna1_at	Azurocidin gene
11	6218	M27783_s_at	ELA2 Elastatse 2, neutrophil
12	6376	M83652_s_at	PFC properdin P factor, complement
13	6308	M57731_s_at	GRO2 GRO2 oncogene
14	1779	M19507_at	MPO myeloperoxidase
15	6185	X64072_s_at	SELL leukocyte adhesion protein beta subunit
16	4082	X05908_at	ANX1 annexin I (lipocortin I)
17	6378	M83667_rna1_s_at	NF-IL6-beta protein mRNA
18	4847	X95735_at	Zyxin
19	5300	L08895_at	MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)
20	1685	M11722_at	Terminal transferase mRNA

Table 9. Classifiers' performance on acute leukemia dataset by cross-validation (number of wrong classified samples in leave-one-out test).

Classification method	Top 5	Top 10	Top 20
FSVM with SVM-RFE	1	0	1
BCT-SVM with F test	2	4	2
BCT-SVM with SVM-RFE	2	1	2

selection. F test and SVM-RFE are used as our gene selection methods combined with multiclass classification methods. In our experiments, three algorithms (FSVM with SVM-RFE, BCT-SVM with SVM-RFE, and BCT-SVM with F test) are tested on three datasets (the real breast cancer data, the small round blue-cell tumors, and the acute leukemia data). The results of these three groups of experiments show that more important genes are selected by FSVM with SVM-RFE, and by these genes selected it shows higher prediction accuracy than the other two algorithms. Compared to some existing multiclass cancer classifiers with gene selection, FSVM based on SVM-RFE also performs very well. Finally, an explanation is provided on the experimental results of this study.

ACKNOWLEDGMENT

This work is supported by China 973 Program under Grant no 2002CB312200 and Center of Bioinformatics Program grant of Harvard Center of Neurodegeneration and Repair, Harvard University, Boston, USA.

REFERENCES

[1] Zhou X, Wang X, Pal R, Ivanov I, Bittner M, Dougherty ER. A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. *Bioinformatics.* 2004;20(17):2918–2927.

[2] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association.* 2002;97(457):77–87.

[3] Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics.* 2002;18(9):1216–1226.

[4] Zhou X, Wang X, Dougherty ER. Multi-class cancer classification using multinomial probit regression with Bayesian variable selection. *IEE Proc of System Biology*. In press.

[5] Zhang HP, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci USA.* 2001;98(12):6730–6735.

[6] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning.* 2002;46(1–3):389–422.

[7] Kim S, Dougherty ER, Barrera J, Chen Y, Bittner ML, Trent JM. Strong feature sets from small samples. *J Comput Biol.* 2002;9(1):127–146.

[8] Zhou X, Wang X, Dougherty ER. Nonlinear-probit gene classification using mutual-information and wavelet based feature selection. *Biological Systems.* 2004;12(3):371–386.

[9] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. *Bioinformatics.* 2003;19(1):90–97.

[10] Zhou X, Wang X, Dougherty ER. Gene selection using logistic regression based on AIC, BIC and MDL criteria. *New Mathematics and Natural Computation.* 2005;1(1):129–145.

[11] Zhou X, Wang X, Dougherty ER. A Bayesian approach to nonlinear probit gene selection and classification. *Franklin Institute.* 2004;341(1-2):137–156.

[12] Zhou X, Wang X, Dougherty ER. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics.* 2003;19(17):2302–2307.

[13] Jornsten R, Yu B. Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics.* 2003;19(9):1100–1109.

[14] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531–537.

[15] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics.* 2000;16(10):906–914.

[16] Mukherjee S, Tamayo P, Mesirov JP, Slonim D, Verri A, Poggio T. Support Vector Machine Classification of Microarray Data. Cambridge, Mass: Massachusetts Institute of Technology; 1999. CBCL Paper 182/AI Memo 1676.

[17] Vapnik VN. *Statistical Learning Theory*. New York, NY: John Wiley & Sons; 1998.

[18] Vapnik VN. *The Nature of Statistical Learning Theory.* 2nd ed. New York, NY: Springer; 2000.

[19] Duda RO, Hart PE, Stork DG. *Pattern Classification.* 2nd ed. New York, NY: John Wiley & Sons; 2001.

[20] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* Belmont, Calif:Wadsworth; 1984.

[21] Abe S, Inoue T. Fuzzy support vector machines for multiclass problems. In: *European Symposium on Artificial Neural Networks Bruges.* Belgium; 2002:113–118.

[22] Inoue T, Abe S. Fuzzy support vector machines for pattern classification. In: *Proceeding of International Joint Conference on Neural Networks.* Washington DC; 2001:1449–1454.

[23] Kreßel UH-G. Pairwise classification and support vector machines. In: Schölkopf B, Burges CJC, Smola AJ, eds. *Advances in Kernel Methods—Support Vector Learning.* Cambridge, Mass:MIT Press; 1999:255–268.

[24] Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med.* 2001;344(8):539–548.

[25] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.* 2001;7(6):673–679.

[26] Yeo G, Poggio T. *Multiclass Classification of SRBCTs*. Cambridge, Mass: Massachusetts Institute of Technology; 2001. CBLC Paper 206/AI Memo 2001-018.

[27] Lee Y, Lin Y, Wahba G. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. *American Statistical Association.* 2004;99(465):67–81.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 172–180 • DOI: 10.1155/JBB.2005.172

RESEARCH ARTICLE

Computational, Integrative, and Comparative Methods for the Elucidation of Genetic Coexpression Networks

Nicole E. Baldwin,[1] Elissa J. Chesler,[2] Stefan Kirov,[3] Michael A. Langston,[1]
Jay R. Snoddy,[3] Robert W. Williams,[2] and Bing Zhang[3]

[1]*Department of Computer Science, The University of Tennessee, Knoxville, TN 37996, USA*
[2]*Department of Anatomy and Neurobiology, The University of Tennessee, Memphis, TN 38163, USA*
[3]*Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

Received 24 June 2004; revised 12 September 2004; accepted 14 September 2004

Gene expression microarray data can be used for the assembly of genetic coexpression network graphs. Using mRNA samples obtained from recombinant inbred *Mus musculus* strains, it is possible to integrate allelic variation with molecular and higher-order phenotypes. The depth of quantitative genetic analysis of microarray data can be vastly enhanced utilizing this mouse resource in combination with powerful computational algorithms, platforms, and data repositories. The resulting network graphs transect many levels of biological scale. This approach is illustrated with the extraction of cliques of putatively coregulated genes and their annotation using gene ontology analysis and *cis*-regulatory element discovery. The causal basis for coregulation is detected through the use of quantitative trait locus mapping.

INTRODUCTION

The purpose of this paper is to describe novel research combining

(i) emergent computational algorithms,

(ii) high performance platforms and implementations,

(iii) complex trait analysis and genetic mapping,

(iv) integrative tools for data repository and exploration.

In this effort we employ huge datasets extracted from a panel of recombinant inbred (RI) strains that were produced by crossing two fully sequenced strains of C57BL/6J and DBA/2J mice [1]. The essential feature of these isogenic RI strains is that they are a genetic mapping panel. They can therefore be used to convert associative networks into causal networks. This is done by finding those polymorphic genes that actually produce natural endogenous variation in gene networks [2]. In this regard, RI strains differ fundamentally from standard inbred strains, knockout strains, transgenic lines and mutants. This approach, termed quantitative trait locus (QTL) mapping, is usually limited to a single continuously distributed trait such as brain weight or neuron number [3], or a behavioral trait such as open-field activity [4]. In this paper, however, we map regulators of entire networks, clusters, and cliques [5].

We employ combinatorial algorithms and graph theory to reduce the high dimensionality of this megavariate data. Advances in clique finding algorithms generate highly distiled gene sets, which we interpret using novel, integrative bioinformatics resources. See Figure 1. Tools of choice include GeneKeyDB [http://genereg.ornl.gov/gkdb], WebQTL [6], and GoTreeMachine (GOTM) [7].

QTL MAPPING

Experimental design

Microarrays provide an extraordinarily efficient tool to obtain very large numbers of quantitative assays from tissue samples. For example, using the Affymetrix M430 arrays one can obtain approximately 45 000 measurements of relative mRNA abundance from a whole tissue

Correspondence and reprint requests to Michael A. Langston, Harvard Center for Neurodegeneration & Repair and Brigham and Women's Hospital, Harvard Medical School, Harvard University, Boston, MA 02115, USA, Email: langston@cs.utk.edu

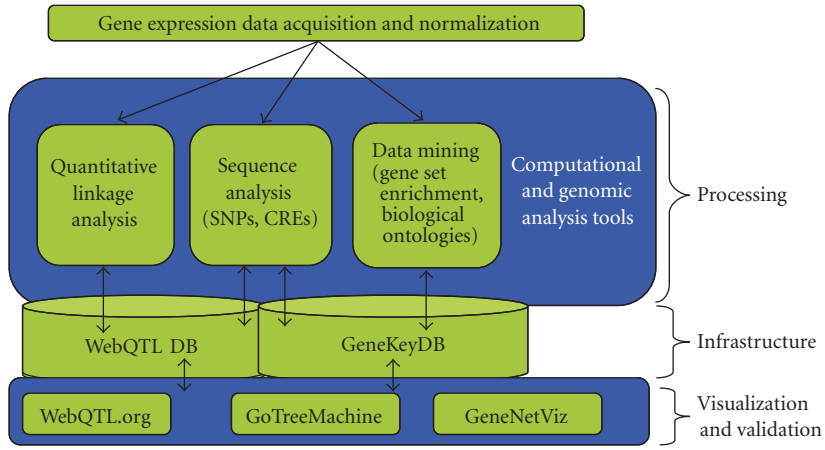


Figure 1. A process overview.

such as the brain or from a single cell population, such as hematopoietic stem cells. In much of our recent work we have used the Affymetrix U74Av2 array to estimate the abundance of 12 422 transcripts from the mouse brain. The design of our experiments is quite simple. We extract mRNA from three litter-mate mice of the same strain and sex, pool the mRNA, and hybridize the sample to the microarray. We do this three times for each strain and sex (independent biological replicates). There is no intentional experimental manipulation of the animals or strains of mice. The essential feature to note in our experimental design is that the isogenic strains of mice that we study are all related yet genetically unique from one another.

RI strains

These related strains of mice collectively form what is called a "mapping panel." The strain set that we use is called the BXD mapping panel because all of the 32 strains originate from the same two original progenitor strains: C57BL/6J (the B strain) and DBA/2J (the D strain). The 32 derivative BXD strains are genetic mosaics of the two parental strains. If one were to pick one of the 32 strains at random and examine a piece of one particular chromosome, there would be an approximately 0.5 probability of that piece having descended down through the generations from the B or the D parental strain. If one looked at the same part of a particular chromosome in all the 32 strains one would end up with a vector of genotypes. For example, the tip of chromosome 1 of BXD strains 11 through 16 might read BDBBDD. Thus there are 2^{32} or 4.29×10^9 possible combinations of these vectors of genotypes. The chromosomes of individual BXD strains actually consist of very long stretches of B-type or D-type chromosomes. The average stretch is almost 50 million base pairs long. The entire set of 32 BXD strains incorporate sufficient recombinations between the parental chromosomes to encode a total of about 2^{11} locations across the mouse genome. This means that in the best case one can only specify locations to about 1.27×10^6 bp. An amount 1.27 million bp will typically contain 17 genes. (Of course, the locations of these recombinations is close to a random Poisson process.)

Unlike other recombinant cross progeny used for QTL analysis, all of the BXD strains are fully inbred. To make these RI strains, full siblings were mated successively for 20 generations to produce each of the 32 strains. This has been an expensive process that has made several strain sets available to the research community by commercial suppliers (The Jackson Laboratory, www.jax.org) or the originating laboratory. Making fully RI strains from a crossbreeding between the C57BL/6J and DBA/2J parental strains has taken about eight years. These strains have been used for over 20 years for the detection of QTLs in a wide range of phenotypes [8]. Additional 45 strains have been generated recently [9].

Finding the genetic regulatory locus

There are numerous genetic polymorphisms (allelic variants) that exist between the two parental strains. As an illustrative example, consider two alleles of a gene coding for a product that is absolutely required to deposit pigmentation in the hair and eye. Further let us assume that these two alleles act as a digital switch: the B allele inherited from C57BL/6J is the active form and the D allele inherited from DBA/2J is the inactive form.

In the D state, the mice are albinos; in the B state, they are normally pigmented. The vector of this phenotype across the strains might look PWWWWP (P, pigmented; W, white) for strains 11 through 16. A simple comparison of this vector of phenotypes to the vector of genotypes on the tip of chromosome 1 (BDBBDD) clearly rules out this location since the vectors do not match particularly well. A vector of genotypes on chromosome 7 at

77 million base pairs (Mb), however, is a perfect fit: BDDDDB. Depending on the coding convention that we use, this will give a correlation either of 1 or of −1. This is the central concept of mapping simple one-gene (monogenic) traits to discover one or more genotype vectors (markers) that have tight quantitative associations with the phenotype vectors across a large mapping panel. Recall however that our particular 32-strain genotype vector only provides enough resolution to get us down to a genetic neighborhood containing about 17 genes. We call this a *genetic locus* (sometimes called a gene locus), although we have to remember that we cannot yet assert which gene in this locus is actually the pigmentation switch.

Up to this point we have considered a trait that can be easily dichotomized. The vast majority of traits in which we are interested, however, are spread continuously over a broad range of values that often approximates a Gaussian distribution. These traits are frequently controled by more than a single genetic locus. Furthermore, environmental factors typically introduce a complementary nongenetic source of variance to a trait measured across a genetically diverse group of individuals. Consider, for example, body weight. This is a classic example of a complex highly variable population trait that is due to a multifactorial admixture of genetic factors, environmental factors, and interactions between genes and environment. Even a trait such as the amount of mRNA expressed in the brains of mice and measured using microarrays is a very complex trait. We refer the interested reader to our previous work [5, 7] for more information on this subject. The abundance of mRNA is influenced by rates of transcription, rates of splicing and degradation, stages of the circadian cycle, and a variety of other environmental factors. Many of these influences on transcript abundance exert their effects via the actions of other genes. QTL mapping of mRNA abundance allows one to detect these genetic sources of variation in gene expression [5, 7, 10, 11].

COMPUTATIONAL METHODS

A clique-centric approach

Current high-throughput molecular assays generate immense numbers of phenotypic values. Billions of individual hypotheses can be tested from a single BXD RI transcriptome profiling experiment. QTL mapping, however, tends to be highly focused on small sets of traits and genes. Many public users of our data resources approach the data with specific questions of particular gene-gene and/or gene-phenotype relationships [12]. These high-dimensional datasets are best understood when the correlated phenotypes are determined and analyzed simultaneously. Data reduction via automated extraction of coregulated gene sets from transcriptome QTL data is a challenge. Given the need to analyze efficiently tens of thousands of genes and traits, it is essential to develop tools to extract and characterize large aggregates of genes, QTLs, and highly variable traits.

There are advantages of placing our work in a graph-theoretic framework. This representation is known to be appropriate for probing and determining the structure of biological networks including the extraction of evolutionarily conserved modules of coexpressed genes. See, for example, [13, 14, 15]. A major computational bottleneck in our efforts to identify sets of putatively coregulated genes is the search for cliques, a classic graph-theoretic problem. Here a gene is denoted by a vertex, and a coexpression value is represented by the weight placed on an edge joining a pair of vertices. Clique is widely known for its application in a variety of combinatorial settings, a great number of which are relevant to computational molecular biology. See, for example, [16]. A considerable amount of effort has been devoted to solving clique efficiently. An excellent survey can be found in [17].

In the context of microarray analysis, our approach can be viewed as a form of clustering. A wealth of clustering approaches has been proposed. See [18, 19, 20, 21, 22] to list just a few. Here the usual goal is to partition vertices into disjoint subsets, so that the genes that correspond to the vertices within each subset display some measure of homogeneity. An advantage clique that holds over most traditional clustering methods is that cliques need not be disjoint. A vertex can reside in more than one (maximum or maximal) clique, just as a gene product can be involved in more than one regulatory network. There are recent clustering techniques, for example those employing factor analysis [23], that do not require exclusive cluster membership for single genes. Unfortunately, these tend to produce biologically uninterpretable factors without the incorporation of prior biological information [24]. Clique makes no such demand. Another advantage of clique is the purity of the categories it generates. There is considerable interest in solving the dense k-subgraph problem [25]. Here the focus is on a cluster's edge density, also referred to as clustering coefficient, curvature, and even cliquishness [26, 27]. In this respect, clique is the "gold standard." A cluster's edge density is maximized with clique by definition.

The inputs to clique are an undirected graph G with n vertices, and a parameter $k \leq n$. The question asked is whether G contains a clique of size k, that is, a subgraph isomorphic to K_k, the complete graph on k vertices. The importance of K_k lies in the fact that each and every pair of its vertices is joined by an edge. Subgraph isomorphism, clique in particular, is $\mathcal{NP}$-complete. From this it follows that there is no known algorithm for deciding clique that runs in time polynomial in the size of the input. One could of course solve clique by generating and checking all $\binom{n}{k}$ candidate solutions. But this brute force approach requires $O(n^k)$ time, and is thus prohibitively slow, even for problem instances of only modest size.

Our methods are employed as illustrated in Figure 2. We will concentrate our discussion on the classic maximum clique problem. Of course we also must handle the related problem of generating all maximal cliques once

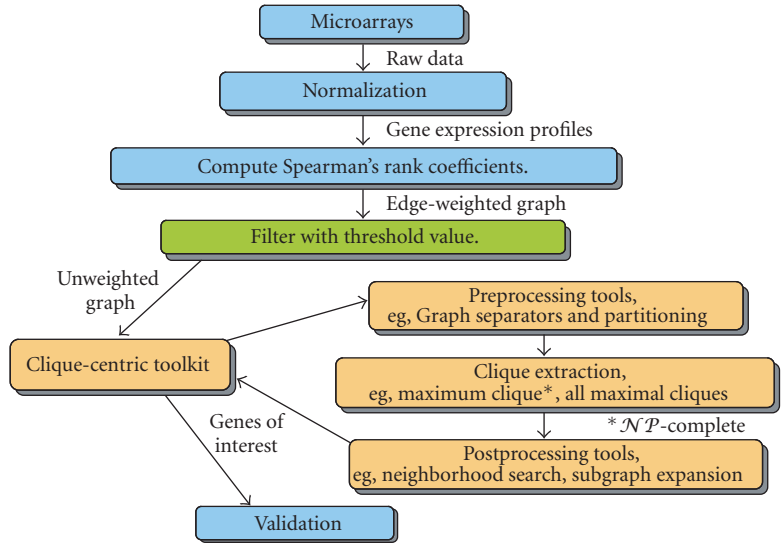


FIGURE 2. The clique-centric toolkit and its use in microarray analysis.

a suitable threshold has been chosen, which is itself often a function of maximum clique size. There are a variety of other issues dealing with preprocessing and postprocessing. Although we do not explicitly deal with them in the present paper, they are for the most part quite easily handled and are dwarfed by the computational complexity of the fundamental clique problem at the heart of our method.

Fixed-parameter tractability

The origins of *fixed-parameter tractability* (FPT) can be traced at least as far back as the work done to show, via the graph minor Theorem, that a variety of parameterized problems are tractable when the relevant input parameter is fixed. See, for example, [28, 29]. Formally, a problem is FPT if it has an algorithm that runs in $O(f(k)n^c)$, where n is the problem size, k is the input parameter, and c is a constant independent of both n and k [30]. Unfortunately, clique is not FPT unless the $\mathcal{W}$ hierarchy collapses. (The $\mathcal{W}$ hierarchy, whose lowest level is FPT, can be viewed as a fixed-parameter analog of the polynomial hierarchy, whose lowest level is $\mathcal{P}$.) Thus we focus instead on clique's complementary dual, the *vertex cover* problem. Consider $\overline{G}$, the complement of G. ($\overline{G}$ has the same vertex set as G, but edges present in G are absent in $\overline{G}$ and vice versa.) As with clique, the inputs to vertex cover are an undirected graph G with n vertices, and a parameter $k \leq n$. The question now asked is whether G contains a set C of k vertices that covers every edge in G, where an edge is said to be covered if either or both of its endpoints are in C. Like clique, vertex cover is $\mathcal{NP}$-complete. Unlike clique, however, vertex cover is also FPT. The crucial observation here is this: a vertex cover of size k in $\overline{G}$ turns out to be exactly the complement of a clique of size $n - k$ in G. Thus, we search for a minimum vertex cover in $\overline{G}$, thereby finding the desired maximum clique in G. Currently, the fastest known vertex cover algorithm runs in $O(1.2852^k + kn)$ time [31]. Contrast this with $O(n^k)$. The requisite exponential growth (assuming $\mathcal{P} \neq \mathcal{NP}$) is therefore reduced to a mere *additive* term.

Kernelization, branching, parallelization, and load balancing

The initial goal is to reduce an arbitrary input instance down to a relatively small computational kernel, then decomposing it so that an efficient, systematic search can be conducted. Attaining a kernel whose size is quadratic in k is relatively straightforward [32]. Ensuring a kernel whose size is linear in k has until recently required much more powerful and considerably slower methods that rely on linear programming relaxation [33, 34].

In [35], we introduced and analyzed a new technique, termed *crown reduction*. A *crown* is an ordered pair (I, H) of subsets of vertices from G that satisfies the following criteria: (1) $I \neq \emptyset$ is an independent set of G, (2) $H = N(I)$, and (3) there exists a matching M on the edges connecting I and H such that all elements of H are matched. H is called the *head* of the crown. The *width* of the crown is $|H|$. This notion is depicted in Figure 3.

Theorem (see [35]). *Any graph G can be decomposed into a crown (I, H) for which H contains a minimum-size vertex cover of G and so that $|H| \leq 3k$. Moreover, the decomposition can be accomplished in $O(n^{5/2})$ time.*

The problem now becomes one of exploring the kernel efficiently. A branching process is carried out using a binary search tree. Internal nodes represent choices; leaves denote candidate solutions. Subtrees spawned off at each

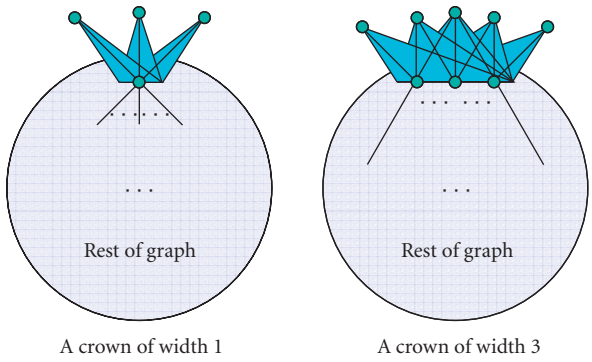


Figure 3. Sample crown decompositions.

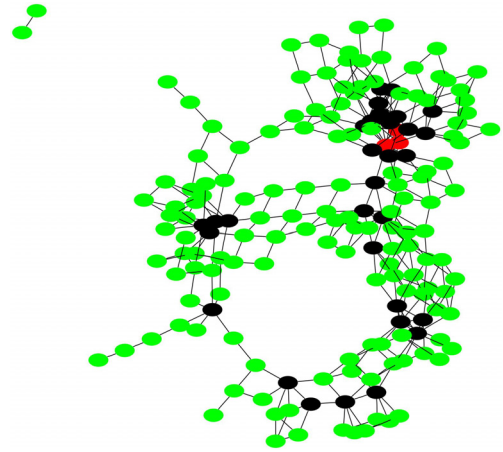


Figure 4. A clique intersection graph for a large microarray dataset.

level can be explored in parallel. The best results have generally been obtained with minimal intervention, in the extreme case launching secure shells (SSHs) [36]. To maintain scalability as datasets grow in size and as more machines are brought on line, some form of dynamic load balancing is generally required. We have implemented such a scheme using sockets and process-independent forking. Results on 32–64 processors in the context of motif discovery are reported in [37]. Large-scale testing using immense genomic and proteomic datasets are reported in [38].

SAMPLE COMPUTATIONAL RESULTS

We are now able to solve real, nonsynthetic instances of clique on graphs whose vertices number in the thousands. (Just imagine a straightforward $O(n^k)$ algorithm on problems of that size!)

To illustrate, we recently solved a problem on *Mus musculus* neurogenetic microarray data with 12 422 vertices (probe sets). With expression values normalized to $[0, 1]$ and the threshold set at 0.5, the clique we returned (via vertex cover) denoted a set of 369 genes that appear experimentally to be coregulated. This took a few days to solve even with our best current methods. Yet solving it at all was probably unthinkable just a short time ago. After iterating across several threshold choices, a value of 0.85 was selected for detailed study. For this graph, G, the maximum clique size is 17. Because it is difficult to visualize G, we employ a clique intersection graph, C_G, as follows. Each maximal clique of size 15 or more in G is represented by a vertex in C_G. An edge connects a pair of vertices in C_G if and only if the intersection of the corresponding cliques in G contains at least 13 members. C_G is depicted in Figure 4, with vertices representing cliques of size 15 (in green), cliques of size 16 (in black), and cliques of size 17 (in red). One rather surprising result is that the gene found most often across large maximal cliques is *Veli3* (aka *Lin7c*). This appears not to be due to some so-called "housekeeping" function, but instead because the relatively unstudied *Veli3* is in fact central to neurological function [39, 40].

CLIQUES OF HIGHLY CORRELATED TRANSCRIPTS AND BEHAVIORAL PHENOTYPES

We can infer that coexpression of genes in mice of common genetic background is due to a shared regulatory mechanism, because the correlation is between trait means from different lines of mice, rather than from within an experimental group. Clique membership alone does not tell us anything about the basis of common genetic regulation. By combining clique data with QTL analysis, the regulatory loci underlying the shared genetic mediation of gene expression can be identified. This allows us to determine the impact of genetic variability in gene expression on other biological processes. Using the aforementioned stringent correlation threshold of 0.85, the most highly connected transcript identified was that of *Veli3*. One maximal clique of seventeen highly associated transcript abundances includes several nuclear proteins. A single principal component of these transcripts accounts for 95% of the total genetic sample variance.

No single QTL can be found for the members of this clique, but a multiple QTL mapping analysis reveals an interacting pair of loci on chromosomes 12 and X, at markers *D12Mit46* (29.163 Mb) and *DXMit117* (110.670 Mb). See Figure 5, which shows the results from a search for pairs of genetic loci that modulate expression of a clique. Chromosomes 12, 19, and X are shown. Likelihood ratio statistics for multiple QTL models are plotted on the pseudocolor scale. The upper left triangle shows fit results for an interaction model, and the lower right triangle shows fit results for a model containing both additive and interaction effects of the two loci. The joint model including markers on 12 and X is significant ($P < .05$) by permutation analysis. A D allele at both loci results in low levels of the phenotype and a B allele at both loci results in a high level of the phenotype. The chromosome 12 locus is the physical location of two clique members: B cell receptor associated protein Bcap29, and myelin transcription

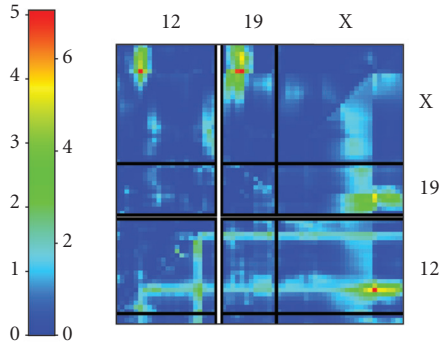


Figure 5. Multiple QTL mapping analysis. In the upper left triangle, a pseudo-color plot shows the likelihood ratio statistic for each two-locus interaction. In the lower right triangle, a likelihood ratio statistic is depicted for the full two-locus model, which fits additive effects for each pair of loci and their interaction. Significance was assessed by genome-wide permutation analysis.

factor 1-like protein, Myt1l. An interesting functional and positional candidate at the Chromosome X locus is integral membrane protein 1, Itm1. While this is not a member of the clique we are analyzing here, it does frequently cooccur along with *Veli3* in many maximal cliques.

In addition to tens of thousands of transcript traits, we have assembled a database of over 600 organismic phenotypes, including many morphometric traits. An understanding of the genetic control of these phenotypes can help explain their evolution. In the present example, we have found the previously mentioned clique to associate with behavioral and metabolic phenotypes. This clique correlates with both midbrain iron levels, and locomotor behavior. Interestingly, one of the clique members, *Gs2na* (GS2 nuclear autoantigen), that, at 46.048 Mb on chromosome 12, is a little too far afield to be a positional QTL candidate gene, is a striatin family member and the negative correlation we observe between clique expression and locomotor behavior is consistent with literature reports of locomotor impairment associated with decreased levels of striatin [41]. At this point research becomes hypothesis driven; indeed, the result of this collaborative analysis is a simple testable hypothesis, extracted from many billions of data relations. We are now in the process of evaluating the hypothesis that genetic variation in iron metabolism influences expression of the *Veli3* clique members in the brain and consequently affects locomotor activity.

INTEGRATIVE GENOMIC DATA MINING

GeneKeyDB

High-throughput, high volume data like these gene expression data from genetically variant mice should be examined in a biological context. The subsets of interesting genes must be analyzed, in part, by using existing information that describe the role these genes play in biological processes. When computing and navigating these data in terms of graphs and networks, we need to have a way to manipulate various kinds of metadata about sets of genes and gene products.

We have developed several such tools for genes and gene products that are discovered from the clique and QTL data analysis. Most gene-centered data resources that are generally available for retrieving metadata about genes are displayed and manipulated in a one-gene-at-a-time format (eg, Entrez Gene). We have developed a lightweight data mining environment that allows the automated integration of various types of data about sets of genes. This environment is called *GeneKeyDB*. This system includes metadata from GenBank, Entrez Gene, Ensembl, and several other well-established biological databases. GeneKeyDB uses a relational database backend to facilitate interactions between tools and data. Among other functions, GeneKeyDB automatically converts the different database identifiers from these different databases. It can, for example, start with GenBank cDNA identifiers, locate the "sequence feature" information from genome sequence data entries, and assist in retrieving sequences for detailed analyses. GeneKeyDB can also obtain various kinds of homologs, functional annotation, or other attributes of genes and gene products. Furthermore, it serves as a repository for results that are created by our computational tools.

We have devised two types of computational analyses that are supported by the underlying GeneKeyDB system.

GoTreeMachine

Gene ontology (GO) produces structured, precisely defined, common, controled vocabulary for describing the roles of genes and gene products [42]. GO has been used frequently in the functional profiling of high-throughput data. We have developed a web-based tool, GOTM, for the analysis and visualization of sets of genes using GO hierarchies [7]. Besides being a stand-alone functional profiling tool, GOTM can work with other computational tools for gene set centered integrated analysis. GOTM has been employed in various ways in this respect. This includes WebQTL's use of GOTM to narrow down candidate gene lists and generate functional profiles for genes in a relevance network or genes correlated to complex phenotypes.

We use ontology analysis to evaluate the functional significance of the cliques found by our graph algorithms, and prioritize the cliques for further study. Figure 6 depicts a clique of size eight that was detected within the gene coexpression network constructed using the microarray data from the RI mouse lines. The five green vertices denote genes that belong to the GO functional category of "DNA binding." The red vertices denote genes that either have no annotation or are annotated as function "unknown." If we randomly pick five genes from all annotated genes on the microarray, the expected frequency of genes in the category DNA binding is only 0.9. The chance

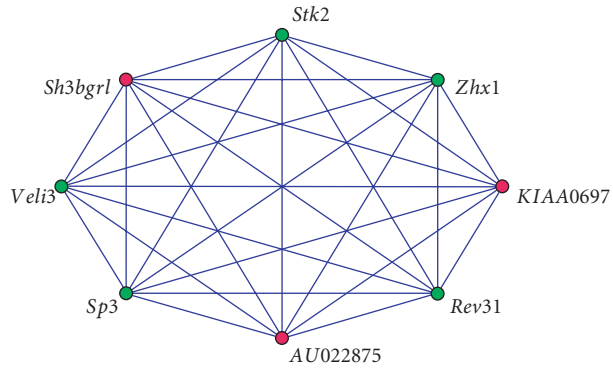


Figure 6. A relevant clique containing *Veli3*.

of finding all five genes in the category DNA binding is $P = .00051$ as calculated by the hypergeometric test implemented in GOTM. Out of the 5227 maximal cliques we generated, ontology analysis has detected a total of 342 of them that are significantly enriched in one or more GO categories ($P < .01$). The clique shown in Figure 6 has a P value less than .001, and is one of several cliques we are studying. Note the presence of the gene *Veli3*.

Batch sequence analysis

We are also deploying integrative methods that attempt to predict *cis*-regulatory elements (CREs) in the upstream regions from sets of genes that are putatively coregulated. These CREs are thought to be the DNA sites to which protein regulatory transcription factors preferentially bind in promoters or enhancers and exert regulatory control of gene expression. We are combining a number of analyses to look at sets of CREs that are found in a subset of genes that seems to show strong coregulation and are consistent with the clique and QTL data.

We have assembled a pipeline, batch sequence analysis (BSA), that can retrieve the sequence data for the target genes and their orthologous counterparts in other chordate organisms. This pipeline carries out a number of processes that enable us to use both coregulated gene sets and phylogenetic footprinting in an integrated pipeline to identify putative CREs. An important advantage of the pipeline is its ability to define the evolutionarily conserved non-coding sequences, which are thought to contain most of the CREs [43]. This should substantially reduce the noise levels. BSA can be carried out in a high throughput, automated process because of the underlying GeneKeyDB infrastructure. BSA is routinely using both multiple Em for motif elicitation (MEME) and motif alignment and search tool (MAST) as part of the sequence analysis, but other motif finding and searching methods are under development. A set of CRE motifs can be found in cliques or other interesting subsets of genes with motif searches (like MEME). We can then use MAST or similar searching tools to take those sets of putative CREs to do a global search for all possible targets in a database that contains promoter sequences from all human, mouse, and rat genes. The latter step could help define new genes that are targets of a gene regulatory network that were not initially identified. The BSA pipeline stores its results in the GeneKeyDB relational database.

SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

Our current work demonstrates the use of clique to extract signal from large genetic correlation matrices. We also employ genome-scale tools to interpret the shared molecular function, biological process, cellular localization, and sequence motifs of clique members. Despite what has been accomplished in the BXD lines, the size of existing RI strain sets limits the power and resolution of this technique. The Complex Trait Consortium plans to expand this set with the development of a 1024 RI strain panel [44]. The creation of this resource will greatly increase the depth of our analysis. The breadth of the analysis can be expanded almost indefinitely. Although the work we have described here has been restricted to the analysis of gene expression microarray phenotypes, any attribute of these strains that can be measured can in principle be incorporated into the genetic correlation matrix. We already have a wealth of data on microscale and macroscale biological phenotypes ranging from cellular responses to behavior. Novel high-throughput molecular phenotypes will greatly expand this collection. To accomodate such vast increases in data dimensionality, we are currently in the process of porting our codes to supercomputers at Oak Ridge National Laboratory (ORNL) (Tennessee, USA). These are difficult tasks indeed, given the many novel features of our algorithms. Great care is required to manage processor and memory resources. Load balancing can be especially problematic [37]. Initial targets include a 256-node SGI Altix and a 256-node Cray X1. In the longer term, we aim to employ the tremendously more powerful machines now under construction and awarded to ORNL in the recent competition to build the nation's next leadership-class computing facility for science. We believe that with our algorithms and these platforms we can solve the problem instances previously considered hopelessly out of reach.

ACKNOWLEDGMENTS

This research has been supported in part by the National Science Foundation under grants CCR–0075792 and CCR–0311500; by the National Institute of Mental Health, National Institute on Drug Abuse, and the National Science Foundation under award P20–MH–62009; by the National Institute on Alcohol Abuse and Alcoholism under INIA grants P01–Da015027, U01–AA013512–02, U01–AA13499, and U24–AA13513; by the Office of Naval Research under grant N00014–01–1–0608;

and by the Department of Energy under contracts DE–AC05–00OR33735 and DE–AC05–4000029264. We wish to thank Drs. Lu Lu and Kenneth Manly for helping develop datasets, analytic tools, and the WebQTL website, www.WebQTL.org. We also wish to express our appreciation to Drs. Mike Fellows and Henry Suters for helping develop fast kernelization alternatives and to Dr. Faisal Abu-Khzam for greatly improved branching implementations.

REFERENCES

[1] Williams RW, Gu J, Qi S, Lu L. The genetic structure of recombinant inbred mice: High-resolution consensus maps for complex trait analysis. *Genome Biol.* 2001;2(11):Research0046.

[2] Airey DC, Shou S, Lu L, Williams RW. Genetic sources of individual differences in the cerebellum. *Cerebellum.* 2002;1(4):233–240.

[3] Peirce JL, Chesler EJ, Williams RW, Lu L. Genetic architecture of the mouse hippocampus: Identification of gene loci with selective regional effects. *Genes Brain Behav.* 2003;2(4):238–252.

[4] Flint J. Analysis of quantitative trait loci that influence animal behavior. *J Neurobiol.* 2003;54(1):46–77.

[5] Chesler EJ, Wang J, Lu L, Qu Y, Manly KF, Williams RW. Genetic correlates of gene expression in recombinant inbred strains: A relational model system to explore neurobehavioral phenotypes. *Neuroinformatics.* 2003;1(4):343–357.

[6] Wang J, Williams RW, Manly KF. WebQTL: web-based complex trait analysis. Neuroinformatics. 2003;1(4):299–308.

[7] Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics.* 2004;5(1):16.

[8] Taylor BA, Wnek C, Kotlus BS, Roemer N, MacTaggart T, Phillips SJ. Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. *Mamm Genome.* 1999;10(4):335–348.

[9] Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* 2004;5(1):7.

[10] Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science.* 2002;296(5568):752–755.

[11] Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature.* 2003;422(6929):297–302.

[12] Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat Neurosci.* 2004;7(5):485–486.

[13] Alon U. Biological networks: the tinkerer as an engineer. *Science.* 2003;301(5641):1866–1867.

[14] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5(2):101–113.

[15] Oltvai ZN, Barabasi AL. Systems biology. Life's complexity pyramid. *Science.* 2002;298(5594):763–764.

[16] Setubal JC, Meidanis J. *Introduction to Computational Molecular Biology.* Boston, Mass: PWS Publishing Company; 1997.

[17] Bomze IM, Budinich M, Pardalos PM, Pelillo M. The maximum clique problem. In: Du D-Z, Pardalos PM, eds. *Handbook of Combinatorial Optimization (Supplement Volume A)*. vol. 4. Boston, Mass: Kluwer Academic Publishers; 1999:1–74.

[18] Bellaachia A, Portnoy D, Chen Y, Elkahloun AG. E-CAST: a data mining algorithm for gene expression data. In: *2nd Workshop on Data Mining in Bioinformatics (BIOKDD 2002).* Alberta, Canada; 2002:49–54.

[19] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. *J Comput Biol.* 2000;7(3-4):559–583.

[20] Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol.* 1999;6(3-4):281–297.

[21] Hansen P, Jaumard B. Cluster analysis and mathematical programming. *Math Program.* 1997;79(1-3):191–215.

[22] Hartuv E, Schmitt A, Lange J, Meier-Ewert S, Lehrachs H, Shamir R. An algorithm for clustering cDNAs for gene expression analysis. In: *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB '99)*. Lyon, France; 1999:188–197.

[23] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA.* 2000;97(18):10101–10106.

[24] Girolami M, Breitling R. Biologically valid linear factor models of gene expression. *Bioinformatics.* 2004;20(17):3021–3033.

[25] Feige U, Peleg D, Kortsarz G. The dense k-subgraph problem. *Algorithmica.* 2001;29:410–421.

[26] Rougemont J, Hingamp P. DNA microarray data and contextual analysis of correlation graphs. *BMC Bioinformatics.* 2003;4(1):15.

[27] Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. *Nature.* 1998;393(6684):440–442.

[28] Fellows MR, Langston MA. Nonconstructive tools for proving polynomial-time decidability. *J ACM.* 1988;35(3):727–739.

[29] Fellows MR, Langston MA. On search, decision and the efficiency of polynomial-time algorithms. *Journal of Computer and Systems Science.* 1994;49:769–779.

[30] Downey RG, Fellows MR. *Parameterized Complexity.* Berlin: Springer; 1999.

[31] Chen J, Kanj IA, Jia W. Vertex cover: further observations and further improvements. *J Algorithms.* 2001;41:280–301.
[32] Buss JF, Goldsmith J. Nondeterminism within $\mathcal{P}$. *SIAM J Comput.* 1993;22(3):560–572.
[33] Khuller S. The vertex cover problem. *SIGACT News.* 2002;33:31–33.
[34] Nemhauser GL, Trotter LE. Vertex packing: Structural properties and algorithms. *Math Program.* 1975;8:232–248.
[35] Abu-Khzam FN, Collins RL, Fellows MR, Langston MA, Suters WH, Symons CT. Kernelization algorithms for the vertex cover problem: Theory and experiments. In: *Proceedings ACM-SIAM Workshop on Algorithm Engineering and Experiments (ALENEX '04).* New Orleans, La; 2004.
[36] Abu-Khzam FN, Langston MA, Shanbhag P. Scalable parallel algorithms for difficult combinatorial problems: A case study in optimization. In: *Proceedings, International Conference on Parallel and Distributed Computing and Systems (PDCS '03).* California; 2003:563–568.
[37] Baldwin NE, Collins RL, Langston MA, Leuze MR, Symons CT, Voy BH. High performance computational tools for motif discovery. In: *Proceedings IEEE International Workshop on High Performance Computational Biology (HiCOMB '04).* Santa Fe, New Mexico; 2004.
[38] Abu-Khzam FN, Langston MA, Shanbhag P, Symons CT. *Scalable Parallel Algorithms for FPT Problems.* Knoxville, Tenn: The University of Tennessee; 2004. Technical Report UT-CS-04-524.
[39] Becamel C, Alonso G, Galeotti N, et al. Synaptic multiprotein complexes associated with 5-HT(2C) receptors: a proteomic approach. *EMBO J.* 2002;21(10):2332–2342.
[40] Butz S, Okamoto M, Sudhof TC. A tripartite protein complex with the potential to couple synaptic vesicle exocytosis to cell adhesion in brain. *Cell.* 1998;94(6):773–782.
[41] Bartoli M, Ternaux JP, Forni C, et al. Down-regulation of striatin, a neuronal calmodulin-binding protein, impairs rat locomotor activity. *J Neurobiol.* 1999;40(2):234–243.
[42] Ashburner M, Ball CA, Blake JA. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet.* 2000;25(1):25–29.
[43] Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet.* 2000;26(2):225–228.
[44] Vogel G. Genetics. Scientists dream of 1001 complex mice. *Science.* 2003;301(5632):456–457.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 181–188 • DOI: 10.1155/JBB.2005.181

RESEARCH ARTICLE

Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases

Nadim W. Alkharouf,[1,2] D. Curtis Jamison,[2] and Benjamin F. Matthews[1]

[1] *Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA*
[2] *School of Computational Sciences, George Mason University, Fairfax, VA 22030, USA*

Received 27 July 2004; revised 26 November 2004; accepted 7 December 2004

Gene expression databases contain a wealth of information, but current data mining tools are limited in their speed and effectiveness in extracting meaningful biological knowledge from them. Online analytical processing (OLAP) can be used as a supplement to cluster analysis for fast and effective data mining of gene expression databases. We used Analysis Services 2000, a product that ships with SQLServer2000, to construct an OLAP cube that was used to mine a time series experiment designed to identify genes associated with resistance of soybean to the soybean cyst nematode, a devastating pest of soybean. The data for these experiments is stored in the soybean genomics and microarray database (SGMD). A number of candidate resistance genes and pathways were found. Compared to traditional cluster analysis of gene expression data, OLAP was more effective and faster in finding biologically meaningful information. OLAP is available from a number of vendors and can work with any relational database management system through OLE DB.

INTRODUCTION

Until recently, data mining required expensive and cumbersome data mining software or a database expert who could accurately translate a request for information into a functional, preferably efficient, query. Database warehouses and online analytical processing (OLAP) offer an attractive and readily available alternative.

As compared to a database, a data warehouse has faster retrieval time, internally consistent data, and a construction that allows users to slice and dice (ie, extract a single item (slice) and compare items in a cross-tabulated table (dice)). The primary difference between a data warehouse and a traditional transaction database lies in the volatility of the data. The information in a transaction database is constantly changing, whereas data in a data warehouse is stable; its information is updated at standard intervals (monthly or weekly). A perfect data warehouse would be updated to add values for the new time period only, without changing values previously stored in the warehouse. Thus, microarray databases can be data warehouses, because the data in them is consistent and stable. Gene expression values in any given experiment remain the same and usually only new data from new experiments is added. Data warehousing software is incorporated in most of the major relational database management systems such as SQLServer2000 and Oracle 9i.

OLAP represents a class of software that enables decision support and reporting based upon a data warehouse [1]. A schematic view of how OLAP software interacts with the data warehouse is shown in Figure 1. OLAP allows for the fast analysis of shared multidimensional information. It is fast because most system responses to users are delivered within 5 seconds, with the simplest analysis taking no more than 1 second and very few taking more than 20 seconds. However, speeds vary by OLAP vendor and system hardware. The key feature of OLAP is that it provides a multidimensional, conceptual view of the data, including full support for hierarchies and multiple hierarchies.

OLAP's underlying structure is the cube [2]. A cube is defined by any number of data dimensions; it is not limited to three; and sometimes an OLAP cube may have fewer than three dimensions. The data dimensions describe an OLAP cube just as width, height, and depth

Correspondence and reprint requests to Benjamin F. Matthews, Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA; E-mail: matthewb@ba.ars.usda.gov

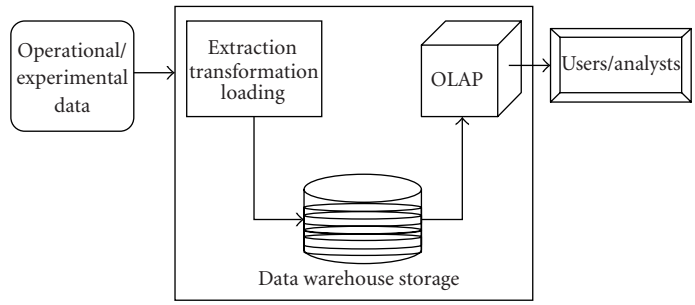


Figure 1. OLAP, cubes and where they fit in a data warehousing solution. OLAP provides efficient and easy-to-use reporting tools and graphical interface, to enable users to mine a data warehouse for hidden information.

describe a geometrical cube. Where it is appropriate, dimensions can be organized into any number of levels (hierarchies).

In relational database systems, OLAP cubes are constructed from a fact table and one or more dimension tables. A fact table is the relational table in the warehouse that stores the detailed values for measures (the thing you are measuring). For example, this could be the values for the relative change in gene expression. The dimension tables however are more abstract, containing only one row for each leaf (lower) member in the fact table. They are used to create summaries and aggregates of the data in the fact table. Ad hoc calculations and statistical analysis can also be achieved, but are vendor specific. Analysis Services 2000 (used here) is capable of such ad hoc calculations on complex data.

The relationship between two dimensions can be modeled using a grid as shown in Table 1. Dimensions are the labels along the axes of the grid and each of the cells is a fact. Facts correspond to the cross product of each dimension of the cube. The data in the cell is a measure, a numerical value. A cube is designed to aggregate, analyze, and find trends in the measures. For example, if the cube describes relative gene induction, the measure is the average relative expression level of a gene under experimental conditions compared to control conditions, and the cube is used to compute this average for the dimensions chosen. In other words, the measure is the number that you would find in the grid cell.

Dimensions are organized into smaller units by using levels where necessary. Levels may also contain other levels, depending on how they are configured in the cube. For example, in Table 1 which represents a two-dimensional cube from our data warehouse designed to identify soybean cyst nematode (SCN) resistance-associated genes in soybean cultivars Peking (P) and Kent (K), the biosamples are considered one level under $K+/K-$ (Kent infected with SCN versus uninfected), which in turn is another level (along with $P+/P-$; Peking infected with SCN versus uninfected) under the dimension probe combination. A fact describes the combination of the various dimensions, for example, probe combination = $P+/P-$, gene = A01A10, time = 6 hours yields a specific fact about the induction of gene A01A10 in $P+/P-$ 6 hours after SCN infection (assuming we added a third dimension of time). This representation is just like the (x, y, z) coordinate system in mathematics. Depending on the way the cube is being used, the fact may show a measure of the induction of a gene at a specific biosample or the result of the t test or some other differential gene expression test.

Table 1. The organization of a cube with two dimensions. In this example, probe combination and genes are dimensions; $P+/P-$, $K+/K-$, biosample 1, biosample 2, A01A10, SSH1B07, D09H12, and B03C02 are levels of the respective dimension. The cells containing various figures are facts. Individual data in the fact cells are the values of the measures. In this example, there are two measures used in the cube, one is the fold induction, the second is the result of the t test (1 significantly induced, -1 significantly suppressed, 0 unchanged).

Fold induction/t test		Probe combination		
		$P+/P-$	$K+/K-$	
			Biosample 1	Biosample 2
Genes	A01A10	1.2	1.5	0.76
		1	-1	1
	SSH1B07	0.34	2.3	-0.98
		0	1	-1
	D09H12	-1.6	1.4	0.03
		-1	1	0
	B03C02	2	1.8	-2.1
		1	1	-1

The meaning of the measure depends on how the cube is defined. The value represents an aggregation for the defined grouping. The measures inside the cube are always numeric. The mathematical operations of count and sum are the primary reason why data warehouses are useful. Calculated measures, such as average, can be calculated from those two basic measures. These are called aggregations. Once dimensions are organized and a cube is being processed, the aggregations are calculated. Generally,

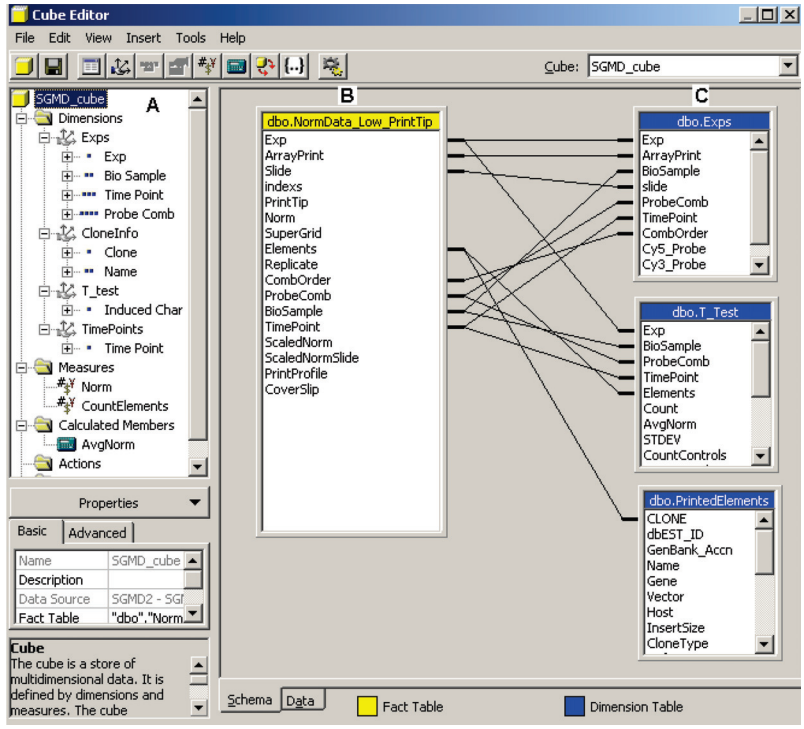


FIGURE 2. A snapshot of a multidimensional cube of gene expression data constructed in Microsoft's Analysis Services 2000 (shipped with SQLServer2000). (A) shows the dimensions of the cube and their associated levels, (B) is the fact table, and (C) shows the dimension tables.

aggregations are calculated immediately after the cube is initially populated or when there is a change in the content of the cube.

OLAP has been used to make some important discoveries in the biomedical field. For instance, Dzeroski et al [3] used OLAP on a database of patients with Y chromosome deletions and found correlations between deletion patterns and patient populations, as well as clinical phenotype severity. OLAP has also been used in the health management field. For example, Silver et al [4] used OLAP to make business decisions that improved operational efficiency of hospitals while maintaining high levels of patient care. Hristovski et al [5] found OLAP to be a suitable data mining tool for public health. However, to the best of our knowledge, OLAP has not been applied to gene expression databases.

We applied OLAP technology to our microarray warehouse, the soybean genomics and microarray database (SGMD) [6], to mine a time-course experiment aiming at discovering genes expressed in soybean roots upon infection by the SCN. SCN is the major pest of soybean and causes an estimated loss of 1$ billion in the United States per year. The discovery of genes expressed under these conditions will provide scientists with information and tools to develop soybean cultivars that are resistant to SCN. Using OLAP we identified numerous candidate genes and associated pathways in a susceptible soybean cultivar (Kent) after infection with SCN [7, 8]. In comparison to traditional gene expression data mining methods, such as k-means and self-organizing maps (SOM) clustering, OLAP performed significantly better at finding candidate genes for further study.

METHODS

Cube construction

We used Analysis Services 2000 (Microsoft, Redmond, Wash), a product that comes with SQLServer2000, to build a multidimensional cube of gene expression experiments conducted over time (Figure 2). Our fact table contained rows of data describing clones and their fold induction at each time point for each biosample and probe combination ($P+/P-, K+/K-$). The measures from this fact table were the normalized log ratio from Lowess print-tip normalization [9], called norm, and the count of unique clones printed (called CountElements). A calculated measure, named AvgNorm, was created to represent the average normalized log ratio from the two measures mentioned above. Four dimensions were created. The first was experiments (exps), which had four levels, exp, biosample, time point, and probe combination. A second dimension, called CloneInfo, had two levels, the clones ID's and their names. The third (t test) and fourth (TimePoints) dimensions had one level each,

induced char (which refers to the results of the t test) and time point, respectively (Figure 2). Cubes are very flexible, new dimensions and measures can be added and removed to customize the data analysis process, that is, the cube can be configured to answer the scientific question at hand.

Microarray data

Gene expression levels of approximately 6000 soybean genes were measured at seven time points after SCN infection [7]. The standard reference design was used for these microarray experiments. The reference (control) sample was RNA extracted from soybean (cultivar Kent) roots which is SCN susceptible, not infected with SCN, and our treatment samples were RNA extracted from Kent cultivar 6 hours, 12 hours, 24 hours, 2 days, 4 days, 6 days, and 8 days after infection with SCN. Reverse labeling of probes was conducted because the two dyes ($Cy3$ and $Cy5$) may not have the same labeling efficiencies and do not have exactly the same correspondence between mRNA concentration and fluorescent intensities. Each gene was printed in triplicate on glass slides. Two replicated slides (one of which is the dye swap) were used for each time point. Two biological samples were also used to account for biological variation and inherent variation in the extraction of mRNA, generating a total of $7 \times 2 \times 2 = 28$ slides and 12 data points for each gene. Self-self hybridized slides were generated for t test analysis. t tests were used to determine differentially expressed genes at each of the time points [9]. Details on slide printing, hybridization, and scanning protocols are described in Alkharouf et al [7]. OLAP was used to produce lists of common significantly induced/suppressed genes at the early (6, 12, and 24 hours), mid (2 and 4 days), and late (6 and 8 days) time points. We used results of the t test to determine significance ($P \leq .05$) and chose a cutoff value of 1.5 fold for extra stringency. In addition, k-means and 2D SOM clustering were applied on the time series data set. k-means was done using J-Express version 2.0 (MolMine; http://www.molmine.com) setting $K = 20$, initialization method to Forgy, and distance metric to Euclidean. SOM was done using the 2D SOM algorithm from GeneSight version 3.5.2 (BioDiscovery; http://www.biodiscovery.com), setting the number of horizontal clusters to 5, the number of vertical clusters to 5, distance metric to Euclidean and clustering by genes only.

RESULTS

OLAP was used to drill down, slice, and dice the time series data and find lists of genes induced and suppressed in each of the specified time intervals (Table 2). OLAP was used to find commonly induced or suppressed genes at two or more time points and in one or more biosamples. OLAP was very quick and efficient in providing those reports. On average OLAP only took 2 to 5 seconds to return a result of a query after the cube was constructed (running on a 1.8 GHz Pentium 4 workstation with 1 GB RAM). This is a fraction of the time needed to produce similar reports from complex SQL queries and multiple-table joins. For instance selecting statistically induced genes common to the 6-, 12-, and 24-hour time points, which requires 3-table joins, took almost 25 seconds to achieve, whereas the same report took only 1 second with OLAP running on the same system.

A common technique for viewing multidimensional output is to view the output as a two-dimensional "slice" of a cube. This is the way the Microsoft SQLServer2000 analysis services display output. This is a simple and informative technique to view the reports in a spreadsheet-like manner. Multidimensional extensions (MDX) can also be used to query cubes instead of using the user interface mentioned above. MDX is a syntax designed for querying multidimensional objects and data and is more flexible than the user interface. It was used to query the cube and obtain the results shown in Table 2. MDX has a similar syntax to SQL, but is designed to work with multidimensional cubes instead of relational tables. The SQLServer2000 analysis services manager has an interface that accepts MDX queries.

The OLAP reports highlighted a number of genes and defense pathways that were triggered in soybean in response to SCN infection (Table 2). These are discussed in detail in [7]. The key findings in the study were that the nematodes elicit the activation of a transcription factor (WRKY) that shuts down a defense pathway known as the salicylic acid inducible pathway, thereby rendering the plants more susceptible to nematode infection.

OLAP found a number of candidate resistance genes that k-means and SOM did not (Table 2), whereas cluster analysis did not reveal any new information that OLAP did not identify by MDX queries. For instance OLAP found trehalose-6-phosphate synthase (TPS) induced at the mid time points, whereas cluster analysis did not. TPS is a key enzyme of sugar metabolism and its induction at the mid time points, where the nematode has formed the syncytium (feeding site), may be an indicator of the parasite's success in utilizing the plants metabolite synthesis apparatus for its own sustenance. Metabolic profiling experiments conducted in collaboration with the Noble foundation also show increased levels of trehalose in Kent 48 hours after infection with SCN (unpublished data). OLAP also found jasmonic acid (JA) inducible genes, such as pathogenesis-related protein PR-6 and chalcone synthase, induced at the early and mid time points whereas cluster analysis did not. The JA signaling pathway is known to be induced in plants after wound damage or parasitic infection [18].

Generally, we found OLAP a lot more powerful for determining genes induced at specific time intervals but not at other time points. This was hard to do using cluster analysis, because the algorithms are designed to group genes with similar profiles, not necessarily to identify

Table 2. Genes found to be induced at different time intervals using OLAP, *k*-means, and SOM clustering. Many of the key candidate resistance genes were identified by OLAP and not cluster analysis, in particular those genes induced at specific time intervals and not others. Cluster analysis did not reveal any other genes that OLAP did not.

Time	GeneID	GeneName	OLAP	*k*-means	SOM	Comments
Induced at all time points	BM139889	Proline-rich glycoproteins	✓	—	—	Cell wall proteins that are found activated during pathogen attack [10] to reinforce the cell wall
	BM107775	Peroxidase	✓	✓	✓	Involved in detoxification and is activated during the hypersensitive response in plants against pathogen attack [11]
	BM139591	Cytochrome P450 monooxygenase	✓	—	—	Photosynthesis-related gene
	BM107779	Photosystem II core proteins	✓	✓	✓	Involved in plant photosynthesis and energy production
	BM107798	4-coumarate-CoA ligase	✓	✓	✓	Involved in phenylpropanoid metabolism and the synthesis of secondary metabolites that are known to be involved in plant defense [12]
	BM108156	Transcription factor WRKY6	✓	✓	✓	Believed to suppress PR-1 genes, thereby inferring susceptibility to pathogen attack in plant species [13]
Induced at the early time points only	CA850582	Trypsin inhibitor proteins	✓	—	—	Proteinase inhibitors
	BM107847	Germin-like protein	✓	—	—	Known to have antimicrobial activity, activated in plants during pathogen infection [14]
	CA851099	Pathogenesis-related protein PR-6	✓	—	—	Proteinase inhibitors known to be induced by jasmonic acid [15]
Induced at the mid time points only	DUP21F10	Trehalose-6-phosphate synthase (TPS)	✓	—	—	Synthesizes trehalose, is thought to be an important regulator of sugar metabolism [16]
	BM108164	Pyrophosphatase	✓	—	—	Metabolism-related gene
	BM108095	Sali3-2 protein	✓	—	—	Induced by aluminum in soybean roots [17]
	BM107806	Chalcone synthase	✓	—	—	Induced by the jasmonic acid signaling pathway [18]
Induced at the late time points only	BM108193	Glutamate dehydrogenase	✓	—	—	Metabolism-related gene
	CA853854	Geranylgeranyl hydrogenase	✓	—	—	Metabolism-related gene
	BM107804	Tyrosine-phosphatase	✓	—	—	Metabolism-related gene
Commonly induced at the early and mid time points	CA850882	Stress-induced gene SAM-22	✓	—	—	A stress-induced PR-10 protein, which is a ribonuclease protein found activated in plants after viral infection [15]
	BM107930	Heat shock protein 70	✓	—	—	Helps new or distorted proteins fold into shape, found induced in a number of plant species after pathogen infection [19]
	BM107821	Lectin-chitin	✓	✓	✓	Cell wall protein

Table 2. Continued

Time	GeneID	GeneName	OLAP	*k*-means	SOM	Comments
Commonly induced at the early and late time points	BM107803	Beta-glucosidase	✓	—	—	Metabolism-related gene
Commonly induced at the mid and late time points	CA852009	Fructose-biphosphate aldolase	✓	—	—	Metabolism-related gene
	BM107809	Sucrose synthase	✓	✓	✓	Metabolism-related gene
	BM108104	ATP-synthase	✓			Metabolism-related gene
	BM108223	Lipoxygenase	✓	✓	✓	Involved in jasmonic acid synthesis and is implicated in plant responses against pathogens [18]
	BM108233	Ubiquitin	✓	—	—	Plays an important role in marking proteins for proteolytic degradation, one of the key events in the systematic defense mechanism of a plant against pathogen invasion [20]
	CA853086	Metallothionein	✓	✓	✓	A member of the aquaporin (AQP) water channel family, induced in rice upon infection with *Magnaporthe grisea* [21]

genes induced uniquely at one time point, but not at others. This explains why none of the genes found uniquely induced at the early, mid, or late time points were identified by cluster analysis (Table 2). Finding these genes is important for the dissection of the metabolic effects of the nematode invasion across time.

In terms of speed, OLAP took approximately 1.2 minutes to generate all the reports summarized in Table 2 and are shown in their entirety on http://psi081.ba.ars.usda.gov/SGMD/Publications/OLAP/. In contrast, it took 5 times longer (approximately 6.5 minutes) to do one of the cluster analysis methods (including the time it takes to export the data from the database to the respective clustering software in the required format). If one were to also measure the time it takes to interpret the OLAP reports versus the clustering results, OLAP would be even at a more advantage point, because it makes the results easier to interpret. Results of the cluster analysis can also be accessed from the web site mentioned above.

DISCUSSION

Gene expression data is valuable for the understanding of gene regulation and biological networks. A main goal of gene expression data analysis is to determine what genes are expressed as a result of a certain cellular state, that is, what genes are expressed in diseased cells that are not expressed in healthy cells. Microarray experiments profile hundreds to thousands of genes at a time generating large data sets that are only getting bigger as more advances in genomics and microtechnologies are made. As these data sets become larger, however, the need for fast and effective database mining tools becomes more obvious and necessary. Data warehouses and OLAP provide tools to construct, populate, view, and access microarray data in an efficient and fast manner. The fundamental unit of OLAP software is the cube, which is a repository of integrated information from the existing data sources.

In our cube design the data sources were the relational tables in SGMD, a gene expression database [6]. Microarray databases are in fact data warehouses because of their consistent and stable data, and little if any modifications to the database model need to be made to use OLAP. OLAP proved to be more efficient than standard relational database queries that rely on time-consuming multitable joins. Although the results obtained from OLAP and these standard SQL queries are the same, the time it takes to execute an OLAP query was found to be 25 times greater than standard SQL queries.

OLAP provides a different view of the data compared to cluster analysis and provides additional insights into the data as shown in Table 2. OLAP identified a number of candidate resistance genes that cluster analysis did not. One reason is the large number of genes of an unknown function that makes such cluster analysis difficult to interpret. OLAP avoids this issue because it allows for the

categorization of genes into categories of known and unknown functions, thereby reducing the complexity of the problem by allowing investigators to analyze genes with a known function first. Another benefit of OLAP is that the values of the clustered elements do not all have to be the same unit, as they are in cluster analysis. This is useful when searching for trends across a heterogeneous data set. In OLAP, you can set any type or number of dimensions to drill your data with, thereby identifying trends that cannot be identified using cluster analysis.

OLAP's main advantage is that it is flexible and can be customized to answer the scientific question at hand if some prior knowledge is known about the data sets, whereas cluster analysis is mostly used as an initial data mining tool with no prior knowledge and is used mainly for grouping genes based on similar expression profiles. The genes that are clustered together however can vary considerably because of the different similarity metrics that are used. Another issue with clustering is that a gene can be characterized in more than one way, while it can belong to only one cluster. OLAP allows scientists, especially those not trained in the computational sciences, to mine their data sets to not only group genes based on their expression profiles but to also ask specific scientific questions such as "give me the genes induced at a certain time point, that is, not induced at all other time points, or the genes induced at time point *A* that are also induced at time points *B*, and *C*," for instance. The answers to these questions can provide valuable insights into the relationships between genes and pathways that cluster analysis cannot answer.

In the case of our data set, for instance, seeking resistance genes induced at specific time points yielded a number of candidate resistance genes and gave us insights into the metabolic changes in soybean when infected with SCN. Thus OLAP is an automation of the manual analysis that most biologists would always perform rather than relying on visually appealing but scientifically uninformative cluster analysis. We are not suggesting that OLAP is better than cluster analysis, but only that the two methods are useful and quite different. We are suggesting however that OLAP can be considered as a supplement or even an alternative to cluster analysis when clustering methods are not suitable to analyze a data set, such as small time-course data sets as ours.

The implementation of OLAP technology to gene expression analysis is not difficult given the right tools. OLAP can be applied to any gene expression database built on any of the major relational database management systems (Oracle, Sybase, MySQL, or even Access), through the use of OLE DB (an industry standard technology for database connectivity). OLAP reports can also be obtained using Excel's (Microsoft, Redmond, Wash) pivot tables, a feature that allows one to cross-tabulate columns in Excel. This might work well for small data sets. OLAP's ability to drill through the data and find common/unique genes given different criteria, along with its flexibility, make it an important data mining tool in gene expression analysis, one that holds great promise in our view.

This study also demonstrates that databases and database applications may not be used solely for the storage and retrieval of expression data but that they can act as tools for doing exploratory data analysis as well. In fact databases can eliminate the need for third-party software, because most of the analysis, even time series analysis, and can be done within the database itself.

REFERENCES

[1] Codd F, Codd SB, Salley CT. *Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate*. San Jose, Calif: Codd EF & Associates;1993. Technical Report.

[2] Gray J, Bosworth A, Layman A, Pirahesh H. *Data cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tabs and Sub-totals.* Washington, DC: Microsoft Corporation; 1995. 95-22. MSR Technical Report.

[3] Dzeroski S, Hristovski D, Peterlin B. Using data mining and OLAP to discover patterns in a database of patients with Y chromosome deletions. *Proc AMIA Symp.* 2000;215–219.

[4] Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ. Case study: how to apply data mining techniques in a healthcare data warehouse. *Healthc Inf Manag.* 2001;15:155–164.

[5] Hristovski D, Rogac M, Markota M. Using data warehousing and OLAP in public health care. *Proc AMIA Symp.* 2000;369–373.

[6] Alkharouf NW, Matthews BF. The soybean genomics and microarray database. *Nucleic Acids Research.* 2004;32:398–400.

[7] Alkharouf N, Chouikha I, Beard H, et al. Expression of soybean genes during invasion of susceptible roots by the soybean cyst nematode. *Mol Plant Microbe Interact*. In press.

[8] Khan R, Alkharouf N, Beard H, et al. Resistance mechanisms in soybean: gene expression profile at an early stage of soybean cyst nematode invasion. *Nematology.* 2004;36(3):241–248.

[9] Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research.* 2002;30(4):e15.

[10] Esquerre-Tugaye M, Campargue C, Mazau D. The response of plant cell wall hydroxyproline-rich glycoproteins to microbial pathogens and their elicitors. In: Datta SK, Muthukrishnan S, eds. *Pathogenesis-Related Proteins in Plants.* Boca Raton, Fla:CRC Press; 1999:157–170.

[11] Low PS, Merida JR. The oxidative burst in plant defense: function and signal transduction. *Physiol Plant.* 1996;96:533–542.

[12] Ryan CA, Jagendorf A. Self defense by plants. *Natl Acad Sci.* 1995;92(10):4075.
[13] Maleck K, Levine A, Eulgem T, et al. The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nature Genetics.* 2000;26(4):403–410.
[14] Schenk PM, Kazan K, Wilson I, et al. Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Natl Acad Sci.* 2000;97(21): 11655–11660.
[15] Van Loon LC. Occurrence and properties of plant pathogenesis-related proteins. In: Datta SK, Muthukrishnan S, eds. *Pathogenesis-Related Proteins in Plants.* Boca Raton, Fla:CRC Press;1999:1–19.
[16] Eastmond PJ, Li Y, Graham IA. Is trehalose-6-phosphate a regulator of sugar metabolism in plants? *Exp Bot.* 2003;54(582):533–537.
[17] Ragland M, Soliman KM. Sali5-4a and sali3-2: two genes induced by aluminum in soybean roots. *Plant Physiology.* 1997;114(3):555–560.
[18] Creelman RA, Mullet JE. Jasmonic acid distribution and action in plants: regulation during development and response to biotic and abiotic stress. *Proc Natl Acad Sci.* 1995;92(10):4114–4119.
[19] Puthoff DP, Nettleton D, Rodermel SR, Baum TJ. *Arabidopsis* gene expression changes during cyst nematode parasitism revealed by statistical analyses of microarray expression profiles. *Plant.* 2003;33(5):911–921.
[20] Kepinski S, Leyser O. Ubiquitination and auxin signaling: a degrading story. *Plant Cell.* 2002;14:81–95.
[21] Kim S, Ahn IP, Lee YH. Analysis of genes expressed during rice-*Magnaporthe grisea* interactions. *Mol Plant Microbe Interact.* 2001;14(11):1340–1346.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 189–197 • DOI: 10.1155/JBB.2005.189

RESEARCH ARTICLE

Cardiovascular Damage in Alzheimer Disease: Autopsy Findings From the Bryan ADRC

Elizabeth H. Corder,[1] John F. Ervin,[2] Evelyn Lockhart,[3] Mari H. Szymanski,[2]
Donald E. Schmechel,[3,4] and Christine M. Hulette[2,3]

[1]*Center for Demographic Studies, Duke University, 2117 Campus Drive, Box 90408, Durham, NC 27708-0408, USA*
[2]*Division of Neurology, Duke University Medical Center, Durham, NC 27710, USA*
[3]*Department of Pathology, Duke University Medical Center, Durham, NC 27710, USA*
[4]*Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA*

Received 7 June 2004; revised 29 November 2004; accepted 7 December 2004

Autopsy information on cardiovascular damage was investigated for pathologically confirmed Alzheimer disease (AD) patients ($n = 84$) and non-AD control patients ($n = 60$). The 51 relevant items were entered into a grade-of-membership model to describe vascular damage in AD. Five latent groups were identified "I: early-onset AD," "II: controls, cancer," "III: controls, extensive atherosclerosis," "IV: late-onset AD, male," and "V: late-onset AD, female." Expectedly, Groups IV and V had elevated *APOE* $\epsilon 4$ frequency. Unexpectedly, there was limited atherosclerosis and frequent myocardial valve and ventricular damage. The findings do not indicate a strong relationship between atherosclerosis and AD, although both are associated with the *APOE* $\epsilon 4$. Instead, autopsy findings of extensive atherosclerosis were associated with possible, not probable or definite AD, and premature death. They are consistent with the hypothesis that brain hypoperfusion contributes to dementia, possibly to AD pathogenesis, and raise the possibility that the *APOE* allele $\epsilon 4$ contributes directly to heart valve and myocardial damage.

INTRODUCTION

It is well known that Alzheimer's disease (AD) is the most common form of senile dementia in the US and Europe. Population studies suggest that 47% of persons over age 85 are affected [1, 2, 3]. The established genetic risk factor is the E4 isoform for the lipid transport molecule apolipoprotein E (*APOE*: gene; ApoE: protein) [4, 5, 6, 7, 8, 9, 10] which is also a risk factor for coronary atherosclerosis [11, 12, 13, 14, 15, 16, 17, 18, 19]. The $\epsilon 4$ allele for *APOE* has sometimes been implicated in vascular dementia (VaD) and stroke [20, 21, 22], the second most common form of senile dementia.

Previous studies have shown decreased smooth muscle actin in brain blood vessels of AD patients when compared to nondemented controls [23]. Possibly, more extensive amyloid deposition in heart and brain vessels determines the 5-fold worse prognosis (8% compared to 40% mortality) within three months following a diagnosis of heart disease or stroke at ages 85+ for $\epsilon 3/4+$ persons, compared to $\epsilon 2/3+$ [24], and largely accounts for reduced $\epsilon 4$, and elevated $\epsilon 2$, frequencies found among centenarians [25].

The role of cardiovascular damage in the development of AD is consistent with a number of existing reports in the literature, although few deal with pathologically confirmed AD and pathologic cardiovascular findings in age-matched samples. Certainly, cardiovascular damage is common among subjects with neuropathologically confirmed AD [26, 27]. Hypertension has been suggested as a risk factor for subsequent AD, normal or low blood pressure at the end stages of the disease [28, 29, 30, 31, 32]. Skoog et al [31] investigated a population cohort and found an association between elevated blood pressure at age 70 years and the development of dementia 10 to 15 years later. They hypothesized that hypertension causes hyalinization of the vessel walls in the brain and hypoperfusion in the deep white matter.

Atherosclerotic disease and silent myocardial infarcts have been associated with cognitive impairment [26, 27]. In the Rotterdam study [33], the extent of atherosclerosis was assessed by ultrasonography of the carotid arteries and by the ratio of ankle to brachial systolic blood pressure. Subjects were scored from 0 to 3, from no to severe, atherosclerosis. The odds of a clinical AD diagnosis

Correspondence and reprint requests to Christine M. Hulette, Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA, E-mail: hulet001@mc.duke.edu

increased 2-fold with the extent of atherosclerosis, 3-fold for VaD.

Autopsy studies are limited, but suggest that cardiovascular disease contributes to the expression of dementia for patients who exhibit Alzheimer neuropathologic changes. Sparks et al [28] showed that significant coronary artery disease was present in 90% (19 of 21) of AD patients undergoing a complete postmortem examination. Patients with peripheral vascular disease, cerebrovascular accidents, and myocardial infarcts had lower antemortem cognitive scores on the minimental state exam [27]. Patients with end-stage renal disease have cognitive impairment thought to be due to multiinfarct dementia [34].

However, Irina et al [35] did not find a correlation of dementia or pathologically confirmed AD with pathologic cardiovascular index (CVI), that is, the extent of atherosclerosis in the brain and periphery combined with evidence of cardiovascular lesions and cardiomegaly. Specifically, the CVI was higher for 103 nondemented subjects compared to 106 demented subjects, 9.2 versus 7.5 out of a possible 15 ($P < .05$). Mean CVI was 5.2 for subjects meeting CERAD criteria for possible AD, 7.3 for definite AD, increasing to 8.5 for vascular and mixed dementia. Thus atherosclerosis was associated with VaD, not the extent of Alzheimer's lesions (associated with the $\epsilon 4$ allele for *APOE*).

To clarify the role of cardiovascular disease in dementia, we investigated 144 subjects prospectively enrolled in the Bryan Alzheimer Disease Research Center Rapid Autopsy Program at Duke University who had had complete body autopsies.

METHODS

The rapid autopsy program

The Rapid Autopsy Program of the Bryan Alzheimer Disease Research Center has been in continuous existence since 1985 [36]. Recruitment, enrollment, and autopsy procedures have been approved by the Institutional Review Board. After receiving informed consent from the patients and their families, both demented and nondemented control donors are enrolled and followed prospectively until death. While the principal purpose of the program is to retrieve and bank human brain tissue for use in research, many donors have consented to complete diagnostic autopsy. At the time of death, consent for autopsy is again obtained according to Duke University Medical Center regulations. Autopsy is performed in the usual fashion with examination of all body organs and cavities. Autopsies are performed in compliance with Centers for Disease Control precautions against the spread of infectious diseases [37, 38].

Data abstraction

APOE genotype (for 46 demented and 38 nondemented subjects) was obtained from existing databases. We abstracted the 144 autopsy records to obtain information on dementia status, cardiovascular disease, medical diagnoses, and organ weights. An Excel spreadsheet was used as the abstract form. Abstraction was done independently by two persons. Inconsistencies were resolved by consensus among the authors. The 51 items are listed in Tables 1–3, respectively.

The statistical approach

Detailed clinical profiles were identified using a statistical technique called grade-of-membership analysis or GoM [39, 40]. Use of univariate approaches would necessarily have low power at this sample size, especially if corrected for multiple comparisons (not needed when all variables are jointly examined). An additional advantage is that each variable can be understood in relation to all the other variables allowing a clinical narrative, something like the process of diagnosis, to be achieved for the identified latent model-based groups.

GoM can be described after first identifying four indices. One is the number of subjects I ($i = 1, 2, \ldots, I$). Here $I = 144$ subjects were identified. The second index is the number of variables J ($j = 1, 2, \ldots, J$). There are $J = 50$ variables each representing one of the clinical variables described above. Our third index is Lj: the set of response levels for the Jth variable.

This leads to the definition of the basic GoM model where the probability that the ith subject has the Ljth level of the Jth variable is defined by a binary variable (ie, $y_{ijl} = 0,1$). The model with these definitions is

$$\text{Prob}\,(y_{ijl} = 1.0) = \sum_k g_{ik}\lambda_{kjl}, \tag{1}$$

where the g_{ik} are convexly constrained scores (ie, $0.0 \le g_{ik} \le 1.0; \sum_k g_{ik} = 1.0$) for subjects and the λ_{kjl} are probabilities that, for the Kth latent group, the Ljth level is found for the Jth variable. The procedure thus uses this expression to identify K profiles representing the pattern of $J \times Lj$ responses found for I subjects.

The parameters g_{ik} and λ_{kjl} are estimated simultaneously using the likelihood function (in its most basic form) [39, 40].

$$L = \prod_i \prod_j \prod_l \left(\sum_k g_{ik} \cdot \lambda_{kjl} \right)^{y_{ijl}}. \tag{2}$$

In the likelihood y_{ijl} is 1.0 if the Ljth level is present and 0.0 if it is not present. GoM models specifying from $K = 3-5$ groups, that is, clinical profiles, were constructed. The significance of adding the $K + 1$ profile was tested as an independent increment in the fit of the model adjusting for the larger number of degrees of freedom in the larger model. Akaike information criterion [41] was calculated as

$$\text{AIC} = -2l(\hat{\theta}) + 2P, \tag{3}$$

Table 1. The clinical variables.

No	Variable	Description
1	Dementia	0 = no, 1 = yes (ie, age at onset listed)
2	Dementia status	0 = normal, 1 = dementia, 2 = normal cognition, 3 = disease control, 4 = early dementia, 6 = Parkinson's disease (PD)
3	Final diagnosis	0 = normal, 1 = AD, 2 = PD, 3 = possible AD, 4 = not listed
4	Age at onset	0 = not demented, 1≤60 years of age, 2 = 60–70, 3 = 70–80, 4 = 80+
5	Duration of dementia	0 = normal, 1 ≤ 5 years, 2 = 5–10, 3 = 10–15, 4 = 15+
6	Age at death	1 ≤ 60 years, 2 = 60–70, 3 = 70–80, 4 = 80+
7	Sex	0 = male, 1 = female
8	Race	0 = white, 1 = black
9	Body mass index	0 = 8–17.9, 1 = 17.9–21.2, 2 = 21.2–25.0, 3 = 25+, 9 = missing
10	Cancer diagnosis	0 = no, 1 = yes
11	Respiratory system infection	0 = no, 1 = yes
12	Urinary system infection	0 = no, 1 = yes
13	Digestive system infection	0 = no, 1 = yes
14	*APOE* genotype	0 = ϵ22, 1 = ϵ23, 2 = ϵ33, 3 = ϵ24, 4 = ϵ34, 5 = ϵ44

where l is the likelihood value and P is the number of estimated parameters. However, for parameters on the boundary, that is, value = 0, only one is penalized. The rationale for subtracting only one for parameters on the boundary is that the distribution for those parameters is $(1/2)X^2$ (*central*). The lowest value of the AIC designates the best model, that is, the model with the best fit and least bias. GoM models specifying either 3, 4, 5, or 6 clinical profiles, that is, $K = 3-6$, had AIC = $-1503, -1647, -1703,$ and -1678, respectively. The 5-group model is reported.

Information on *APOE* genotype was not used to construct the groups used to clinically characterize the subjects. One option in the likelihood is to separate calculations for "internal" (here, clinical) and "external" (here, *APOE* genotype) variables. For internal variables, MLE of g_{ik} and λ_{kjl} are generated and the information in internal variables is used to define the K groups. For external variables the likelihood is evaluated (and MLE of λ_{kjl}; generated) but the information is not used to redefine the K groups, that is, the likelihood calculations for likelihood equations involving the g_{ik} are disabled for external variables so that the g_{ik}, and the definition of the K groups, is not changed.

RESULTS

Overview

Five model-based groups best represented the autopsy information on diagnoses, cardiovascular disease, and organ weights. They are labeled I, II, III, IV, and V ordered according to increasing age at the time of death. Each group is defined by the probabilities of response for the many variables, akin to the frequencies found in the sample as a whole. Tables 4–6 describe the profiles in terms of diagnoses, cardiovascular damage, and organ weights, respectively. Not all the variables described in Tables 1–3 are shown in the tables of results.

The size of the groups was similar (Table 4). The sizes are the summed memberships of individuals in the respective groups either partial, that is, fractional, or complete, that is, contributing size one to the sum, depending on the extent of resemblance of the individual to the group. Group I was the largest group ($n = 36.4$) and Group II was the smallest ($n = 21.3$). The prevalence of demented and nondemented subjects, that is, sum of memberships, in each model-based group is shown in Table 7, as well as the distribution of *APOE* frequencies across the model-based groups.

Dementia diagnoses

Next observe that the groups were either demented or not demented (Table 4): Groups I, IV, and V had 100% probability of dementia while Groups II and III were not demented. There was high probability that dementia was specifically due to AD: 81% for Group I, 73% for Group IV, and 72% for Group V. Otherwise, no explanatory diagnosis was given for the dementia for these groups. Groups II and III had no chance of an AD diagnosis. However, possible AD was sometimes found in the nondemented groups, 9% for Group II and 43% for Group III (5–10 years younger than Groups IV and V). Group II had a 22% chance of being a so-called "disease control" and Group III had a 10% chance of having Parkinson's disease. The dementia status and final diagnoses variables are described in Table 2.

Onset age, sex, and dementia duration

As the age at the onset of dementia for Group I was usually before age 65, Group I represents early-onset AD. Groups IV (male) and V (female) represent late-

Table 2. The cardiovascular variables.

No	Variable	Description
15	Pericardial cavity fluid	0 ≤ 10 mL, 1 = 11–50, 2 = 51–100, 3≥100, 9 = missing
16	Right pleural cavity fluid	0 ≤ 10 mL, 1 = 11–50, 2 = 51–100, 3 ≥ 100, 9 = missing
17	Left pleural cavity fluid	0 ≤ 50 mL, 1 = 51–200, 2 = 201–500, 3 ≥ 500, 9 = missing
18	Peritonial fluid	0 ≤ 50 mL, 1 = 51–200, 2 = 201–500, 3 ≥ 500, 9 = missing
19	Aorta atherosclerosis	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
20	Right coronary artery atherosclerosis	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
21	Right coronary artery narrowing	0 = normal, 1 = 5%–25%, 2 = 26%–75%, 3 ≥ 75%, 9 = missing
22	Right coronary artery calcification	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
23	Left main coronary artery atherosclerosis	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
24	Left main coronary artery narrowing	0 = normal, 1 = 5%–25%, 2 = 26%–75%, 3 ≥ 75%, 9 = missing
25	Left main coronary artery calcification	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
26	Circumflex branch atherosclerosis	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
27	Circumflex branch narrowing	0 = normal, 1 = 5%–25%, 2 = 26%–75%, 3 ≥ 75%, 9 = missing
28	Circumflex branch calcification	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
29	Anterior descending branch atherosclerosis	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
30	Anterior descending branch narrowing	0 = normal, 1 = 5%–25%, 2 = 26%–75%, 3 ≥ 75%, 9 = missing
31	Anterior descending branch calcification	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
32	Right atrial cavity dilation	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
33	Right atrial wall thickness	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
34	Right ventricular cavity dilation	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
35	Right ventricular wall thickness	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
36	Left atrial cavity dilation	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
37	Left atrial wall thickness	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
38	Right ventricular cavity dilation	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
39	Right ventricular wall thickness	0 = normal, 1 = mild, 2 = moderate, 3 = severe, 9 = missing
40	Aortic valve	0 = normal, 1 = abnormal, 9 = missing
41	Tricuspid valve	0 = normal, 1 = abnormal, 9 = missing
42	Pulmonic valve	0 = normal, 1 = abnormal, 9 = missing
43	Mitral valve	0 = normal, 1 = abnormal, 9 = missing
44	Ventricular myocardium	0 = normal, 1 = abnormal, 9 = missing

Table 3. The quartiles for organ weights.

No	Variable	Description
45	Heart	0 ≤ 295 gm, 1 = 295–351, 2 = 351–412, 3 = 412+, 9 = missing
46	Lungs (mean)	0 ≤ 316 gm, 1 = 316–443, 2 = 443–585, 3 = 585+, 9 = missing
47	Liver	0 ≤ 880 gm, 1 = 880–1100, 2 = 1100–1140, 3 = 1140+, 9 = missing
48	Kidneys (mean)	0 ≤ 99 gm, 1 = 99–125, 2 = 125–153, 3 = 153+, 9 = missing
49	Spleen	0 ≤ 75 gm, 1 = 75–113, 2 = 113–170, 3 = 170+, 9 = missing
50	Adrenals (mean)	0 ≤ 6.95 gm, 1 = 6.95–8.1, 2 = 8.1–10.2, 3 = 10.2+, 9 = missing
51	Thyroid	0 ≤ 9.25 gm, 1 = 9.25–14, 2 = 14–20, 3 = 20+, 9 = missing

onset AD: the mean age at onset was around 70 years of age for both groups, marginally earlier for female Group V compared to male Group IV. Nonetheless Group V had longer disease duration and an older age at the time of death. The 2% of the sample that was black was concentrated in the control Groups II and III (not shown).

BMI and nonneurologic diagnoses

Early-onset dementia (I) was associated with extremely low body mass index (BMI). Women with late-onset dementia and long dementia duration (V) also had very BMI. Cancer was a common diagnosis for Group II, absent for Group III. Men with late onset dementia also frequently had cancer (IV), absent for female Group V.

Table 4. Clinical variable frequencies for each group. Each model-based group is defined by the probabilities of being demented & probabilities of response for the other variables, that is, the model λ parameters.

Variable	Response	Group				
		$N = 36.4$	$N = 21.3$	$N = 27.2$	$N = 27.5$	$N = 31.5$
		I	II	III	IV	V
Demented	Yes (%)	100	0	0	100	100
Age at onset	< 60 years	81	—	—	7	0
	60–70	15	—	—	35	53
	70–80	4	—	—	46	34
	80+	0	—	—	12	13
Duration of dementia	< 5 years	12	—	—	5	0
	5–10	38	—	—	62	0
	10–15	43	—	—	33	37
	15+	7	—	—	0	63
Age at death	< 60	20	20	0	0	0
	60–70	73	0	12	0	0
	70–80	7	73	68	31	12
	80+	0	7	20	69	88
Sex	Female	66	20	35	0	100
BMI (km/m^2)	8–17.9	67	0	0	4	23
	17.9–21.2	11	0	0	37	54
	21.2–25.0	13	41	17	51	14
	25+	10	59	83	9	9
Cancer diagnosis	Yes	24	84	0	60	0
Respiratory infection	Yes	69	70	82	57	100
Urinary tract infection	Yes	16	16	27	29	44
Digestive tract infection	Yes	20	0	22	30	54

Respiratory and urinary infections were common for each group especially for demented women (V). Digestive tract infections were not found for Group II (cancer).

Cardiovascular disease

There was little evidence of cardiovascular disease for the early-onset relatively young Group I. Control Group II with cancer uniquely often had pulmonary effusions, ascites, and a moderately dilated right ventricle. Groups II–V had minimal amounts of pericardial cavity fluid, absent for Group I (not shown).

Control Group III without cancer had severe atherosclerosis with narrowing and calcification in the aorta and each of the major coronary vessels—right coronary artery, left main coronary artery, circumflex branch, and anterior descending branch. The extent of atherosclerosis was usually similar for each vessel. Table 5 represents the average over all the coronary vessels. Group III with extensive atherosclerosis often also had moderate atrial dilation and moderately thickened ventricular myocardium (not shown).

Atherosclerosis was unexpectedly less extensive for demented Groups IV and V, limited for Group II, and absent for Group I.

Aortic and mitral valve damage and evidence of ischemic damage to the left ventricular myocardium were common for male late-onset dementia Group IV, and also female Group V. Both of the late-onset dementia groups and Group III had a moderately dilated right atrium.

Organ weights

Generally speaking, the organ weights shown in Table 6 paralleled the BMI results shown in Table 4. Nonetheless there were some interesting departures: organ weights were preserved for Group I having the lowest BMI. The low heart weight for this relatively young group is consistent with limited heart damage, as indicated by Table 4. Lung weight was highest for Group II, which often had pulmonary edema and pleural effusions. Group III, most affected by atherosclerosis, had the highest heart weight. Compared to Group III, organ weights and BMI were lower for Group IV. Very low BMI and weight for most organs was found for late-onset dementia Group V, females with long dementia duration. Notably, both late-onset dementia groups had low thyroid weight compared to the early-onset and control groups.

Table 5. Cardiovascular damage for each group. The model-based groups are defined by the probabilities response for the variables sometimes represented on a semiquantitative scale: "+ + +" denotes severe, while "++," "+," and "−" denote moderate, mild, and the absence of lesions; "+ + /−" denotes mixture of moderate and absent; "+/−" denotes mixture of mild and absent; "+ + /+" denotes a mixture of moderate and mild; "+ + +/−" denotes a mixture of severe and absent. *means the average of right coronary artery, left main coronary artery, circumflex branch, and anterior descending branch.

Outcome	Location	Group				
		I	II	III	IV	V
Fluid	Peritoneal cavity	−	+ + +/−	+/−	−	−
	Pleural cavity	−	+ + +	+/−	−	−
Coronary artery*	Atherosclerosis	−	+/−	+ + +	++	+
	Narrowing	−	+/−	+ + +	++	+
	Calcification	−	+/−	+ + +	++	+
Aorta	Atherosclerosis	+/−	++	+ + +	++	+ + /+
Dilation	Right atrium	−	+/−	+ + /−	++	++
	Right ventricle	+/−	+ + /−	+/−	+	+
	Left atrium	−	+/−	+/−	+/−	+/−
	Left ventricle	+/−	+	+	−	+/−
Damage	Aortic valve	—	20	—	100	33
	Mitral valve	14	32	—	100	45
	Pulmonic valve	—	—	—	37	14
	Tricuspid valve	—	19	3	64	17
	Ventricular myocardium	—	—	—	49	46

APOE genotype

APOE genotype data was available on a subset of 84 subjects, 46 with dementia and 38 controls. Although this information was not used to predict the groups, individuals carrying the $\epsilon 4$ allele were more common in the dementia groups. The summed memberships of individuals of each genotype are shown in Table 7. As individuals, the study subjects who exactly resembled a single profile contributed one to the size of the relevant profile and zero to the other profiles. Otherwise, the subject contributed a total of one to the sizes of the relevant profiles depending on the extent of resemblance. In contrast to results, Tables 4 – 6 that predict frequencies for persons exactly like the group, Table 7 demonstrates that as individuals there was overlap of demented and nondemented subjects in the groups, not surprisingly given the many variables used to construct the groups, frequent comorbidity at advanced ages, and differences from individual to individual.

DISCUSSION

We investigated cardiovascular damage found for 84 demented and 60 nondemented subjects enrolled in the Bryan ADRC Rapid Autopsy Program. The subjects could be represented by five distinct latent groups based on detailed pathologic information. The late-onset AD groups, both male (IV) and female (V), had frequent heart valve damage, evidence of ischemic damage to the left ventricular myocardium, low BMI, and low organ weights notably including the thyroid gland. They did not have extensive atherosclerosis compared to control subjects without cancer (III), many of whom had possible AD. In particular, the female group (V) having long AD duration had little atherosclerosis. Control subjects with cancer (II) had little atherosclerosis or valve damage. Instead, pulmonary edema and ascites were common. The early-onset AD group (I) had little cardiovascular damage, with normal organ weights despite low body weight.

The finding of mitral and aortic valve damage, and evidence of ischemic damage to the left ventricular myocardium for the AD groups, especially among men, is interesting. It is consistent with the hypothesis that brain hypoperfusion and microthrombi may contribute to the evolution of AD pathology or to the expression of dementia at an earlier stage in AD pathogenesis [40].

However, the AD groups (IV and V) were the oldest groups. Thus valve and myocardial damage might simply be age related and less rapidly fatal than extensive coronary atherosclerosis. Alternatively, the oldest cohorts in the sample may have valve damage resulting from rheumatic fever not treated with antibiotics. Assuming that mitral valve damage is related to rheumatic fever in the oldest subjects does not, however, rule out the possibility that it contributes to the expression of dementia by decreasing brain perfusion and, possibly, contributing to brain pathology.

In addition, the lack of an age-matched control group requires comment. Since aging is itself a risk factor for valvular disease, an age-matched control group would be

Table 6. Organ weights for each group. The model-based groups are defined by the probabilities response for the variables, here represented on a semiquantitative scale: "+ + ++" denotes the highest quartile of weight, while "+ + +," "++," and "+" denote the respectively lower quartiles.

Organ	Group				
	I	II	III	IV	V
Heart	+	+++	++++	+++	++
Lungs	++	++++	+++	++	+
Liver	+++	++++	++++	+++	+
Kidneys	+++	+++	++++	+++	+
Spleen	++	++++	++++	++	+
Adrenals	+++	++++	+++	+++	+
Thyroid	+++	++++	++++	+++	+

Table 7. Distribution of subjects across the groups, subdivided according to dementia status and *APOE* genotype; $\epsilon 2/2$ and $\epsilon/2/3$ were grouped with $\epsilon 3/3$; $\epsilon 2/4$ and $\epsilon 4/4$ were grouped with $\epsilon 3/4$. The sizes are the sums of the model g_{ik} parameters, that is, the memberships of individuals in the groups, whole or partial.

Subjects		Group					
		I	II	III	IV	V	All
Demented		30.9	2.3	8.5	19.7	22.6	84
Not demented		5.5	19.0	18.7	7.8	9.0	60
$\epsilon 3/3+$	Demented	4.0	0	2.0	3.9	5.0	15
	Normal	2.4	9.1	8.6	2.2	3.6	26
$\epsilon 4/4$	Demented	12.3	1.0	3.2	8.2	6.2	31
	Normal	1.6	3.5	3.8	1.5	1.5	11
Total		36.4	21.3	27.2	27.5	31.5	144

required to fully examine the hypothesis that hypoperfusion contributes to dementia. Unfortunately, this was not possible in this small-human-population-based study. Rigorous analysis must await examination of a larger cohort.

Despite the caveats, the results on valve and myocardial damage are striking and the given frequent finding of the $\epsilon 4$ allele for *APOE* in the AD groups raises the possibility that ApoE directly damages these structures as it has been demonstrated to damage blood vessels [42, 43]. Muscle actin in the arterioles is replaced by amyloid for $\epsilon 4/4+$ subjects to a much greater extent than for $\epsilon 3/3+$ subjects. Meyer et al [44] ($n = 36$) found hypertrophy of the left ventricle (uncommon in the Bryan ADRC sample) in 9 of 19 (47%) $\epsilon 3/4+$ AD patients and only 1 of 11 (9%) $\epsilon 3/3+$ patients ($\chi^2 2 = 3.8$, df $= 1$, $P = 0.05$). There were no statistically significant differences in the presence of stenotic changes or calcification in aortic or mitral valvulae in this small sample. The study results tend to suggest that the 5-fold worse prognosis following a diagnosis of heart disease or stroke at ages 85+ for $\epsilon 4+$ persons may be due to impaired cerebrovascular function due to amyloid deposition [25]. Data presented here would also suggest that cardiovascular malfunction may be a contributing factor.

The study does not support the notion that extensive atherosclerosis is a risk factor for definite AD. Instead, subjects with extensive atherosclerosis died before a diagnosis of probable or definite AD could be made. The relatively low $\epsilon 4$ frequency for control subjects with extensive atherosclerosis was unexpected since the allele carries a modestly increased risk of coronary atherosclerosis [13, 14, 15, 16, 17, 44]. There is also a small ecologic association between $\epsilon 4$ frequency and population rates of myocardial infarction in middle age [45, 46]. Nevertheless the findings from this study are consistent with the lack of risk for heart disease and stroke for $\epsilon 4$ found in the Kungsholmen Project for cohort age of 75 and older [25].

Atherosclerosis may possibly at least partially reverse itself during the clinical progression of AD as weight is lost and food intake diminished. This explanation is supported to some extent by the fact that possible AD was common in the nondemented group with high heart weight and extensive atherosclerosis. So that atherosclerosis might be a common concomitant of early or preclinical AD but may not be found at the time of death many years later.

A strong feature of the study is that comparisons were made based on pathologic features and pathologically confirmed diagnoses. The data analytic approach was helpful in resolving the many items of information into a tractable number of distinct groups consistent with clinical experience, despite the relatively small sample

size. For example, heart weight was highest for the group also having the most extensive athersosclerosis and lung weight was highest for the group also having pulmonary effusions and ascites.

In summary, extensive coronary atherosclerosis at autopsy was associated with death at earlier ages and limited AD pathology. Pathologically confirmed AD was not associated with extensive coronary atherosclerosis. Surprisingly, it was associated with mitral and aortic valve damage and damage to the ventricular myocardium.

ACKNOWLEDGMENTS

We thank the study subjects and their families who made the project possible. *Hai Huang MD and Yi-Ping Pan MD were responsible for data abstraction.* Financial support for the study was provided by Grants P50 AG05128 and R01AG07198 from the National Institute on Aging (USA), Glaxo Smith-Kline, and numerous small donations from patients and family members to the Joseph and Kathleen Bryan Alzheimer Disease Research Center.

REFERENCES

[1] Evans DA, Funkenstein HH, Albert MS, et al. Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported. *JAMA*. 1989;262(18):2551–2556.

[2] Silver MH, Jilinskaia E, Perls TT. Cognitive functional status of age-confirmed centenarians in a population-based study. *J Gerontol B Psychol Sci Soc Sci*. 2001;56(3):P134–P140.

[3] Seshadri S, Wolf PA, Beiser A, et al. Lifetime risk of dementia and Alzheimer's disease The impact of mortality on risk estimates in the Framingham study. *Neurology*. 1997;49(6):1498–1504.

[4] Strittmatter WJ, Saunders AM, Schmechel D, et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Nat Acad Sci USA*. 1993;90(5):1977–1981.

[5] Saunders AM, Strittmatter WJ, Schmechel D, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*. 1993;43(8):1467–1472.

[6] Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 1993;261(5123):921–923.

[7] Lucotte G, Visvikis S, Leininger-Muler B, et al. Association of apolipoprotein E allele epsilon 4 with late-onset sporadic Alzheimer's disease. *Am J Med Genet*. 1994;54(3):286–288.

[8] Lucotte G, Aouizerate A, Gerard N, Turpin JC, Landais P. Allele doses of apolipoprotein E type epsilon 4 in sporadic late-onset Alzheimer's disease. *Am J Med Genet*. 1995;60(6):566–569.

[9] Lucotte G, Turpin JC, Landais P. Apolipoprotein E-epsilon 4 allele doses in late-onset Alzheimer's disease. *Ann Neurol*. 1994;36(4):681–682.

[10] Schmechel DE, Saunders AM, Strittmatter WJ, et al. Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset Alzheimer disease. *Proc Nat Acad Sci USA*. 1993;90(20):9649–9653.

[11] Cumming AM, Robertson FW. Polymorphism at the apoprotein-E locus in relation to risk of coronary disease. *Clin Genet*. 1984;25(4):310–313.

[12] Davignon J, Gregg RE, Sing CF. Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis*. 1988;8(1):1–21.

[13] Laakso M, Kesaniemi A, Kervinen K, Jauhiainen M, Pyorala K. Relation of coronary heart disease and apolipoprotein E phenotype in patients with non-insulin dependent diabetes. *BMJ*. 1991;303(6811):1159–1162.

[14] van Bockxmeer FM, Mamotte CD. Apolipoprotein epsilon 4 homozygosity in young men with coronary heart disease. *Lancet*. 1992;340(8824):879–880.

[15] Eichner JE, Kuller LH, Orchard TJ, et al. Relation of apolipoprotein E phenotype to myocardial infarction and mortality from coronary artery disease. *Am J Cardiol*. 1993;71(2):160–165.

[16] Stengard JH, Zerba KE, Pekkanen J, Ehnholm C, Nissinen A, Sing CF. Apolipoprotein E polymorphism predicts death from coronary heart disease in a longitudinal study of elderly Finnish men. *Circulation*. 1995;91(2):265–269.

[17] Hallman DM, Boerwinkle E, Saha N, et al. The apolipoprotein E polymorphism: a comparison of allele frequencies and effects in nine populations. *Am J Hum Genet*. 1991;49(2):338–349.

[18] Wilson PW. Established risk factors and coronary artery disease: the Framingham study. *Am J Hypertens*. 1994;7(pt 2):7S–12S.

[19] Gerdes LU, Gerdes C, Kervinen K, et al. The apolipoprotein epsilon 4 allele determines prognosis and the effect on prognosis of simvastatin in survivors of myocardial infarction: a substudy of the Scandinavian simvastatin survival study. *Circulation*. 2000;101(12):1366–1371.

[20] Slooter AJ, Tang MX, van Duijn CM, et al. Apolipoprotein E epsilon 4 and the risk of dementia with stroke. A population-based investigation. *JAMA*. 1997;277(10):818–821.

[21] Goldstein LB, Vitek MP, Dawson H, Bullman S. Expression of the apolipoprotein E gene does not affect motor recovery after sensorimotor cortex injury in the mouse. *Neuroscience*. 2000;99(4):705–710.

[22] Pasquier F, Henon H, Leys D. Risk factors and mechanisms of post-stroke dementia. *Rev Neurol (Paris)*. 1999;155(9):749–753.

[23] Ervin JF, Pannell C, Szymanski M, Welsh-Bohmer K, Schmechel DE, Hulette CM. Vascular smooth muscle actin is reduced in Alzheimer disease brain:

a quantitative analysis. *J Neuropathol Exp Neurol*. 2004;63(7):735–741.

[24] Corder EH, Basun H, Fratiglioni L, et al. Inherited frailty. ApoE alleles determine survival after a diagnosis of heart disease or stroke at ages 85+. *Ann N Y Acad Sci*. 2000;908:295–298.

[25] Schachter F, Faure-Delanef L, Guenot F, et al. Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet*. 1994;6(1):29–32.

[26] Aronson MK, Ooi WL, Morgenstern H, et al. Women, myocardial infarction, and dementia in the very old. *Neurology*. 1990;40(7):1102–1106.

[27] Breteler MM, Claus JJ, Grobbee DE, Hofman A. Cardiovascular disease and distribution of cognitive function in elderly people: the Rotterdam study. *BMJ*. 1994;308(6944):1604–1608.

[28] Sparks DL, Scheff SW, Liu H, Landers TM, Coyne CM, Hunsaker JC 3rd. Increased incidence of neurofibrillary tangles (NFT) in non-demented individuals with hypertension. *J Neurol Sci*. 1995;131(2):162–169.

[29] Prince M, Cullen M, Mann A. Risk factors for Alzheimer's disease and dementia: a case-control study based on the MRC elderly hypertension trial. *Neurology*. 1994;44(1):97–104.

[30] Launer LJ, Masaki K, Petrovitch H, Foley D, Havlik RJ. The association between midlife blood pressure levels and late-life cognitive function. The Honolulu-Asia aging study. *JAMA*. 1995;274(23):1846–1851.

[31] Skoog I, Lernfelt B, Landahl S, et al. 15-year longitudinal study of blood pressure and dementia. *Lancet*. 1996;347(9009):1141–1145.

[32] Guo Z, Viitanen M, Winblad B, Fratiglioni L. Low blood pressure and incidence of dementia in a very old sample: dependent on initial cognition. *J Am Geriatr Soc*. 1999;47(6):723–726.

[33] Hofman A, Ott A, Breteler MM, et al. Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam study. *Lancet*. 1997;349(9046):151–154.

[34] Lass P, Buscombe JR, Harber M, Davenport A, Hilson AJ. Cognitive impairment in patients with renal failure is associated with multiple-infarct dementia. *Clin Nucl Med*. 1999;24(8):561–565.

[35] Irina A, Seppo H, Arto M, Paavo R Sr, Hilkka S. β-amyloid load is not influenced by the severity of cardiovascular disease in aged and demented patients. *Stroke*. 1999;30:613–618.

[36] Hulette CM, Welsh-Bohmer KA, Crain B, Szymanski MH, Sinclaire NO, Roses AD. Rapid brain autopsy. The Joseph and Kathleen Bryan Alzheimer's Disease Research Center experience. *Arch Pathol Lab Med*. 1997;121(6):615–618.

[37] Brown P, Wolff A, Gajdusek DC. A simple and effective method for inactivating virus infectivity in formalin-fixed tissue samples from patients with Creutzfeldt-Jakob disease. *Neurology*. 1990;40(6):887–890.

[38] Centers for Disease Control. Agent summary statement for human immunodeficiency virus and report on laboratory-acquired infection with human immunodeficiency virus. *MMWR*. 1988;37:S5.

[39] Manton KG, Woodbury MA, Tolley HD. *Statistical Applications Using Fuzzy Sets*. New York, NY:John Wiley & Sons; 1994.

[40] Manton KG, Stallard E, Corder LS. The dynamics of dimensions of age-related disability 1982 to 1994 in the US elderly population. *J Gerontol A Biol Sci Med Sci*. 1998;53(1):B59–B70.

[41] Akaike H. Information theory and an extension of the maximum likelihood principle. In: *Proc 2nd International Symposium on Information Theory*. Budapest, Hungary:Akademiai Kiado; 1973:267–281.

[42] Hulette CM, Welsh-Bohmer KA, Murray MG, Saunders AM, Mash DC, McIntyre LM. Neuropathological and neuropsychological changes in "normal" aging: evidence for preclinical Alzheimer disease in cognitively normal individuals. *J Neuropathol Exp Neurol*. 1998;57(12):1168–1174.

[43] Kosunen O, Talasniemi S, Lehtovirta M, et al. Relation of coronary atherosclerosis and apolipoprotein E genotypes in Alzheimer patients. *Stroke*. 1995;26(5):743–748.

[44] Meyer JS, Rauch G, Rauch RA, Haque A. Risk factors for cerebral hypoperfusion, mild cognitive impairment, and dementia. *Neurobiol Aging*. 2000;21(2):161–169.

[45] Otto CM, Lind BK, Kitzman DW, Gersh BJ, Siscovick DS. Association of aortic-valve sclerosis with cardiovascular mortality and morbidity in the elderly. *N Engl J Med*. 1999;341(3):142–147.

[46] Stengard JH, Pekkanen J, Sulkava R, Ehnholm C, Erkinjuntti T, Nissinen A. Apolipoprotein E polymorphism, Alzheimer's disease and vascular dementia among elderly Finnish men. *Acta Neurol Scand*. 1995;92(4):297–298.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 198–214 • DOI: 10.1155/JBB.2005.198

RESEARCH ARTICLE

Metabolite Fingerprinting in Transgenic *Nicotiana tabacum* Altered by the *Escherichia coli* Glutamate Dehydrogenase Gene

R. Mungur,[1,6] A. D. M. Glass,[2,3] D. B. Goodenow,[4] and D. A. Lightfoot[1,3,5]

[1]*Department of Molecular and Medical Biochemistry, Southern Illinois University, Carbondale, IL 62901, USA*
[2]*Department of Botany, University of British Columbia, Vancouver, Canada V6T 1Z4*
[3]*Department of Plant, Biology, Southern Illinois University, Carbondale, IL 62901, USA*
[4]*Phenomenome Discoveries Inc. 941 University Drive, Saskatoon, Canada S7N 0K2*
[5]*Department of Plant, Soil and Agricultural Systems, Southern Illinois University, Carbondale, IL 62901, USA*
[6]*Max-Planck Institute of Molecular Plant Physiology, 14476 Golm, Potsdam, Germany*

Received 14 April 2004; revised 11 June 2004; accepted 27 July 2004

With about 200 000 phytochemicals in existence, identifying those of biomedical significance is a mammoth task. In the postgenomic era, relating metabolite fingerprints, abundances, and profiles to genotype is also a large task. Ion analysis using Fourier transformed ion cyclotron resonance mass spectrometry (FT-ICR-MS) may provide a high-throughput approach to measure genotype dependency of the inferred metabolome if reproducible techniques can be established. Ion profile inferred metabolite fingerprints are coproducts. We used FT-ICR-MS-derived ion analysis to examine *gdhA* (glutamate dehydrogenase (GDH; EC 1.4.1.1)) transgenic *Nicotiana tabacum* (tobacco) carrying out altered glutamate, amino acid, and carbon metabolisms, that fundamentally alter plant productivity. Cause and effect between *gdhA* expression, glutamate metabolism, and plant phenotypes was analyzed by $^{13}NH_4^+$ labeling of amino acid fractions, and by FT-ICR-MS analysis of metabolites. The *gdhA* transgenic plants increased ^{13}N labeling of glutamate and glutamine significantly. FT-ICR-MS detected 2 012 ions reproducible in 2 to 4 ionization protocols. There were 283 ions in roots and 98 ions in leaves that appeared to significantly change abundance due to the measured GDH activity. About 58% percent of ions could not be used to infer a corresponding metabolite. From the 42% of ions that inferred known metabolites we found that certain amino acids, organic acids, and sugars increased and some fatty acids decreased. The transgene caused increased ammonium assimilation and detectable ion variation. Thirty-two compounds with biomedical significance were altered in abundance by GDH including 9 known carcinogens and 14 potential drugs. Therefore, the GDH transgene may lead to new uses for crops like tobacco.

INTRODUCTION

Due to improvements in mass spectrometry (MS), the methods of metabolite analysis are becoming fast, reliable, sensitive, and automated [1] with broad applications to biological phenomena [2, 3, 4]. A range of analytical techniques can be used with complex biological samples. However, the development of ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) have provided robust techniques that can be widely applied [1]. Electron impact quadrupole MS is also evolving toward a robust technology for metabolite analysis [2, 3, 4]. Libraries of compound identities have been developed at a mass accuracy of 10 ppm (about 0.01 d), often by MS-MS fragmentation. In contrast, the mass accuracy of full-scan MS in a Fourier transformed ion cyclotron resonance mass spectrometer (FT-ICR-MS) format provides for mass accuracy to 1 ppm (about 0.001–0.0001 d) if the ion cyclotron is not filled [5, 6, 7]. The greater potential for mass accuracy is derived from the longer path length that allows for separation of a larger number of compounds, protein fragments, or DNA molecules per analysis.

However, with FT-ICR-MS the techniques for robust identification of ions, the methods for inference of the underlying metabolites, the supporting databases, and the methods for quantification are at an earlier stage of development and are less well known than for other MS formats [8]. The abundance of specific ions in total infusion

Correspondence and reprint requests to D. A. Lightfoot, Department of Molecular and Medical Biochemistry, Southern Illinois University, Carbondale, IL 62901, USA, E-mail: ga4082@siu.edu

mass spectra is the result of the combined ion suppression effects of all other components, pH and salinity of the solution, flow rate, tip opening, and electrospray current [9]. Small effects that alter the overall matrix composition may have large effects on total mass spectra. Therefore, although ion fingerprinting by FT-ICR-MS is a valuable tool for detecting subtle effects for mutant classification [10], exact masses alone may not be sufficient to identify specific compounds in more complex comparisons.

Post-genomic research that aims to determine gene function(s) and relationships among pathways and products will require more tools for metabolite analyses [1]. While multiparallel analyses of mRNA and protein abundance provide indirect information on the biochemical function of genes, metabolic analysis can provide direct information on instantiations [4]. Biological function is the sum of gene interactions and metabolic network interactions; both are affected by environment and genetics [11]. Many changes in mutants and transgenic organisms are cryptic, silent, or unpredictable [12, 13, 14, 15, 16]. Metabolite analysis, particularly metabolite fingerprinting and metabolomics, can detect cryptic changes and link unpredictable phenotypes to their biochemistry [4, 16]. Both metabolomics and metabolite profiling can provide information on how the central metabolites regulate cellular metabolism [11].

Glutamate dehydrogenases (GDH; EC 1.4.1.1 and EC 1.4.1.2) catalyze the reversible amination of alpha-ketoglutarate to form glutamate. In plants, they are not expected to assimilate ammonium because the enzyme is located in the mitochondria, is homo-octameric in structure, and has a high Km for substrates compared to glutamine synthetase (EC 6.4.2.1). The effects of genetic modification of nitrogen metabolism via the bacterial glutamate dehydrogenase (homo-hexameric GDH; EC 1.4.1.1) on plant growth and metabolism were not as expected [12, 13, 14, 15]. In the greenhouse and growth chamber herbicide tolerance is provided, biomass increase is increased, and water deficit resistance is increased [12, 13, 14, 15]. In the field, over three consecutive years, relative yield increase was caused by GDH [12]. An overall increase in the concentration of sugars, amino acids, and ammonium ions occurs within the cell [12, 13, 14]. A biochemical alteration may cause this effect, related to increased production of glutamate in one intracellular compartment, the cytoplasm. Increased total carbohydrate and amino acid compositions show that both carbon and nitrogen metabolism are altered in *gdhA* plants [13].

Reported here are the detected ion inferred metabolic fingerprint and changes in ion peak size inferred metabolite abundance among tobacco roots and leaves in plants transgenic for GDH compared to nontransgenic plants. The extent of glutamate synthesis was measured by ^{13}N labeling. These data illustrate the use of FT-ICR-MS as a tool to analyze transgenic plants and to identify chemicals with biomedical significance.

Table 1. Labeling of the glutamate pool via absorption of $^{13}NH_4^+$ in intact roots of transgenic plants not treated with 1 mM MSX. Tracer exposure was for 15 minutes. Incorporation is expressed as a percentage of the label input plus or minus the range detected among 3 individual plant replicates and three measurement replicates.

	BAR	GDH10	GUS
Glu	9.5 ± 1.5	21.3 ± 4.9	9.4 ± 1.4
Gln	32.9 ± 4.3	41.7 ± 3.8	32.7 ± 4.2
NH_4^+	57.5 ± 5.7	36.3 ± 3.5	57.3 ± 6.0

Table 2. Labeling of the glutamate pool via absorption of $^{13}NH_4^+$ in intact roots of transgenic plants treated with 1 mM MSX for 2 hours before feeding. Tracer exposure was for 15 minutes. Incorporation is expressed as a percentage of the label input plus or minus the range detected among 3 individual plant replicates and three measurement replicates.

	BAR	GDH10	GUS
Glu	1.3 ± 0.5	3.2 ± 0.6	1.3 + 0.5
Gln	1.5 ± 0.6	1.9 ± 0.4	1.5 + 0.6
NH_4^+	97.2 ± 0.4	94.9 ± 1.0	97.1 + 0.3

RESULTS

Production of homozygous lines for biochemical evaluations

We had previously generated r2 seed from a series of independently regenerated plants that showed a range of GDH activity of 2–25 μmol $min^{-1}mg^{-1}$ protein [12]. Each line was an independent transformant, with genetic architecture consistent with one or two copies of the *gdhA* transgene [15]. The mRNA abundance and GDH activity were correlated via Northern hybridization. The mRNA was of high abundance for the GDH10 that produced between 20 and 23 μmol $min^{-1}mg^{-1}$ protein GDH activity. GDH10 line was selected for further analyses compared to vector and nontransgenic controls.

Analysis of glutamate fraction labeling

For comparison of glutamate fraction labeling we selected GDH10, GUS, and BAR transgenic tobacco lines because only GDH is expected to be resistant to methionine sulfoximine (MSX), an inhibitor of photorespiratory ammonium assimilation. Comparisons (Tables 1, 2, 3, and 4) did show organ specific differences.

In the roots, labeling of the fraction containing ^{13}N-glutamate (from $^{13}NH_4^+$ administered during a 15-minute period) was increased 2.2 fold in GDH10 compared to BAR and GUS plants as a result of the introduced GDH activity, representing 21% (dpm/dpm) of the $^{13}NH_3$ applied (Table 1). Treatment with MSX, an inhibitor of glutamine synthetase, reduced glutamate fraction labeling 7 fold among the GDH, GUS, and BAR transgenics suggesting the GS/GOGAT cycle accounts for 86% of

Table 3. Labeling of amino acid fractions in leaves fed $^{13}NH_4^+$ through the petiole after 15 minutes and held in nutrient solution. Entire leaves were cut from 3 replicates of tobacco plants that were 6 weeks old grown in soil in a 16/8 walk in growth room at 26°C with light at about 500 microEinsteins. Incorporation is expressed as a percentage of the label input plus or minus the range detected among 3 individual plant replicates and three measurement replicates.

	GUS	GDH10	BAR
Glu	29.7 + 4.0	18.6 + 2.7	23.4 + 3.1
Gln	18.8 + 6.0	12.0 + 1.1	16.3 + 3.3
NH_4^+	51.3 + 4.3	69.5 + 1.8	51.3 + 2.5

Table 4. Labeling of the glutamate pool via absorption of $^{13}NH_4^+$ in entire leaves of transgenic plants treated with 1 mM MSX for 1.5 hours before feeding. Leaf petioles were recut under water. Tracer exposure was for 15 minutes. Incorporation is expressed as a percentage of the label input plus or minus the range detected among 3 individual plant replicates and three measurement replicates.

	GUS	GDH10	BAR
Glu	2.9 ± 1.3	7.0 ± 0.6	2.8 ± 0.7
Gln	1.7 ± 0.1	8.5 ± 1.9	1.5 ± 0.9
NH_4^+	95.3 ± 2.3	84.0 ± 2.4	95.0 ± 4.1

the labeling in the absence of MSX. However, in MSX-inhibited GDH10 roots, glutamate labeling remained 2.2 fold higher than GUS and BAR roots (Table 2). Therefore, GDH was not inhibited by MSX. As expected BAR did not inactivate MSX.

In leaves, both glutamate fraction labeling and total labeling were decreased by 1.2 to 1.5 fold in GDH10 compared to GUS and BAR control plants (Table 3). The decrease was not significant in this experiment or experiments with leaf discs (data not shown). However, glutamate fraction labeling in presence of MSX was decreased 10 fold in GUS and BAR plants but only 2.6 fold in the GDH10 (Table 4). In addition, glutamine fraction labeling in presence of MSX was decreased 10 fold in GUS and BAR plants but only 0.6 fold in the GDH10. Therefore, in MSX-inhibited leaves; GDH10 assimilated 5 fold more ^{13}N than GUS control and BAR plants. The GDH10 line, in the presence of MSX, also left less $^{13}NH_4^+$ unincorporated (84% compared to 95%, Table 4) reflecting the contribution of the *gdhA* gene in NH_4^+ assimilation in MSX-inhibited leaves. The glutamine labeling in MSX-treated GDH10 leaves was not related to incomplete inhibition of GS since the same degree of labeling was observed in 1 cm^3 leaf discs floating in labeling solution (data not shown).

In GDH10 there was 2–3 fold more label in the glutamate fraction of both MSX-treated leaves (7.0% of the absorbed $^{13}NH_4^+$, Table 4) and roots (3.2%, Table 2) than BAR leaves (2.9%, Table 4) and roots (1.3%, Table 2). However, in non-MSX-treated GDH10 transgenic leaves, compared to the roots, the very high activity of GS, the larger pool sizes of glutamate and the greater flux through pathways involving glutamate may have resulted in less labeling by ^{13}N (Tables 1 and 3). The amount of label in the glutamate and glutamine fractions that could be attributed to GDH activity was modest in roots, about 2.3%. ((3.2−1.3) + (1.9−1.5)). However, in leaves, labeling was significantly greater, about 11.9% ((7.0−2.9) + (8.5−1.7)).

Analysis of ion fingerprints and profiles

Experiments with tobacco [13, 14, 15] and corn [17, 18] had indicated that the total soluble amino acid, ammonium, and carbohydrate contents of GDH transgenic plants were each increased. The transgenic seedlings were shown to reproduce this phenotype (Table 5(a)). Ions were separated and characterized to infer the detectable metabolite complement using four ionization protocols for FT-ICR-MS. There were 2 012 ions detected within 2–4 ionization protocols (unique ions and isotope ions were removed). Regardless of genotype, ion fingerprints of leaves and roots differed significantly judged by FT-ICR-MS. Qualitative differences (compounds only detected in one organ) approached 23% (462/2012). Ion masses were validated by internal calibration with compounds of known mass and concentrations. Among the ions common in roots and leaves, apparent quantitative differences were in the majority 60% (929/1550). Quantitative differences were inferred from peak areas and validated by internal calibration. However, many factors can interfere with peak detection so that the estimates of differences in quantity are not unequivocal and some may be erroneous. Within that context some of the data observed were consistent with known organ-specific metabolisms in plants and some were not.

The metabolites we putatively inferred from ion masses that were altered in abundance by GDH activity are depicted in Figures 1, 2, 3, 4, and listed in Table 5. The majority of the metabolites increased or decreased in leaves and/or roots by less than 10 fold. Between 5% and 14% of detectable metabolites were altered in abundance. This portion of the database can be examined at http://www.siu.edu/~pbgc/metabolite-profiles/GDH/.

In leaves, 98 (5%) of the ions detected were changed in abundance between GDH and non-GDH plants. Only 91 empirical formulas could be inferred because seven were equivocal. Forty-one matched the formulas and predicted masses of the ions of compounds found in the databases we searched. The masses of the remaining fifty unidentified metabolite ions are available at http://www.siu.edu/~pbgc/metabolite-profiles/GDH but not discussed further here for brevity. The 41 putatively identified compounds were categorized as follows: 11 amino acids, 2 sugars, 8 fatty acids, 6 compounds of special nitrogen metabolism, 2 nucleic acid derivatives, 1 TCA cycle intermediate, 1 stress-related compound,

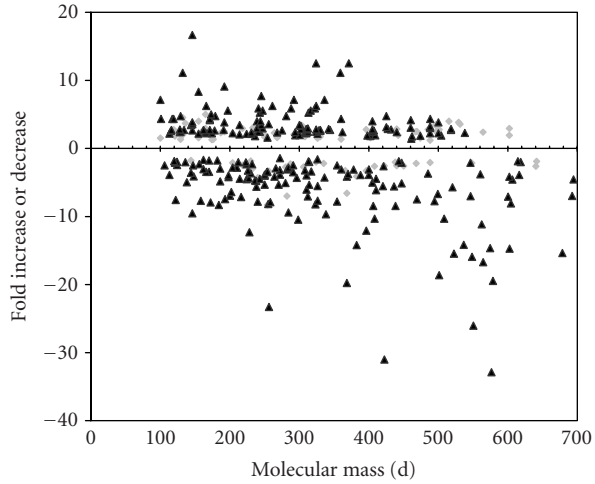


Figure 1. Distribution of metabolites (judged by mass) altered in relative abundance in leaves and roots. Grey diamonds are leaf metabolites and black triangles represent metabolites altered in roots.

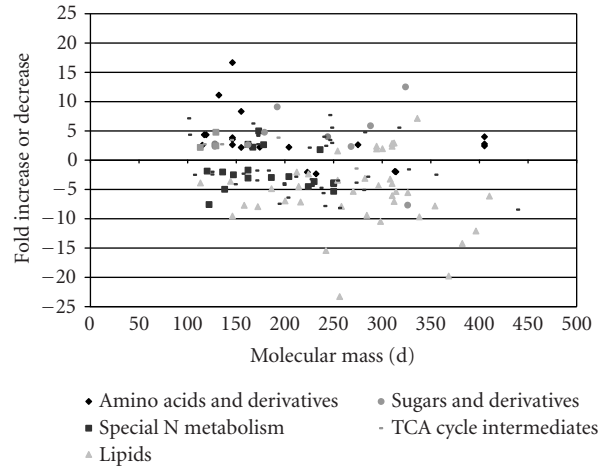


Figure 3. Scatter plot distribution of all classes of metabolites identified in root extracts.

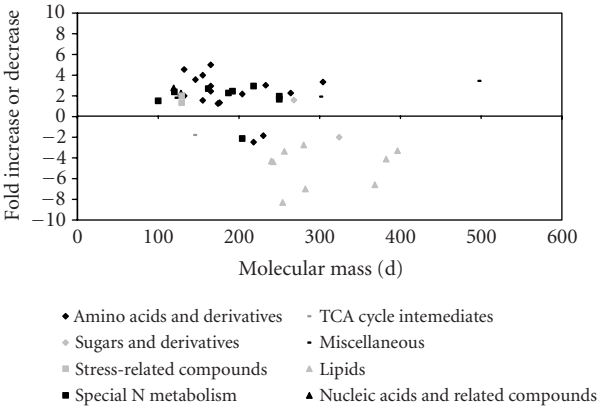


Figure 2. Scatter plot distribution of all classes of metabolites identified in leaf extracts.

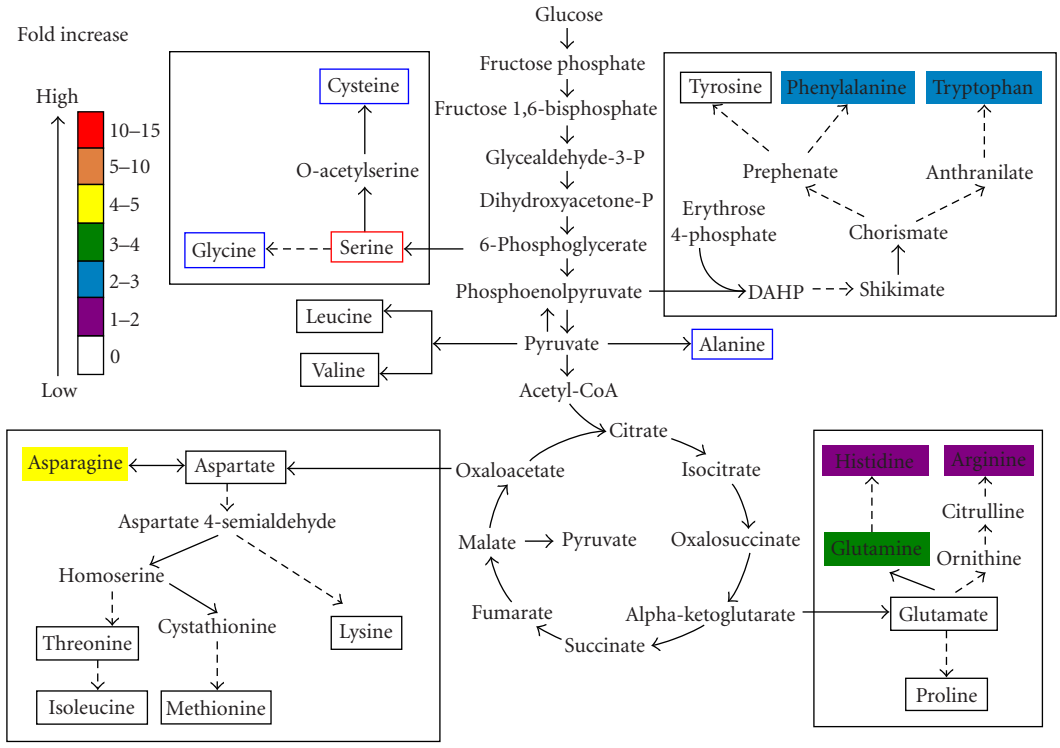
and 10 miscellaneous metabolites not of those classes (Figure 4, Table 5). Not all of these compounds are common metabolites. Some are compounds not previously detected in plant cells, possibly reflecting the animal and microbial fauna present on tobacco samples. Some identified compounds were not previously detected in vivo possibly reflecting ionization artifacts. However, for brevity hereafter the *metabolite putatively inferred from a detected ion* will be referred to as just the *metabolite*.

In roots, there were 283 ions (14%) that changed in abundance among the 2 012 ion species repeatedly detected. Only 268 empirical formulas could be inferred. Database searches putatively identified only 117 of the 283 changed metabolites (Figure 3). Masses of the unidentified ions are available at http://www.siu.edu/~pbgc/metabolite-profiles/GDH but not reported further here for brevity. Among the 117 altered metabolites, there were 14 amino acids, 6 sugars, 34 fatty acids, 15 compounds of special nitrogen metabolism, 2 nucleic acid derivatives, 4 TCA cycle intermediates, 2 stress-related compounds, and 40 metabolites not of those classes (Figure 4, Table 5). Judged by the correspondence of ion mass estimates, 90% of the compounds that changed in abundance in leaves also increased or decreased in the same way in roots. Only three metabolites were altered so that the increase in one organ was accompanied by decrease in the other organ (63 and 75, 86 and 87, and 60 and 67, Table 5).

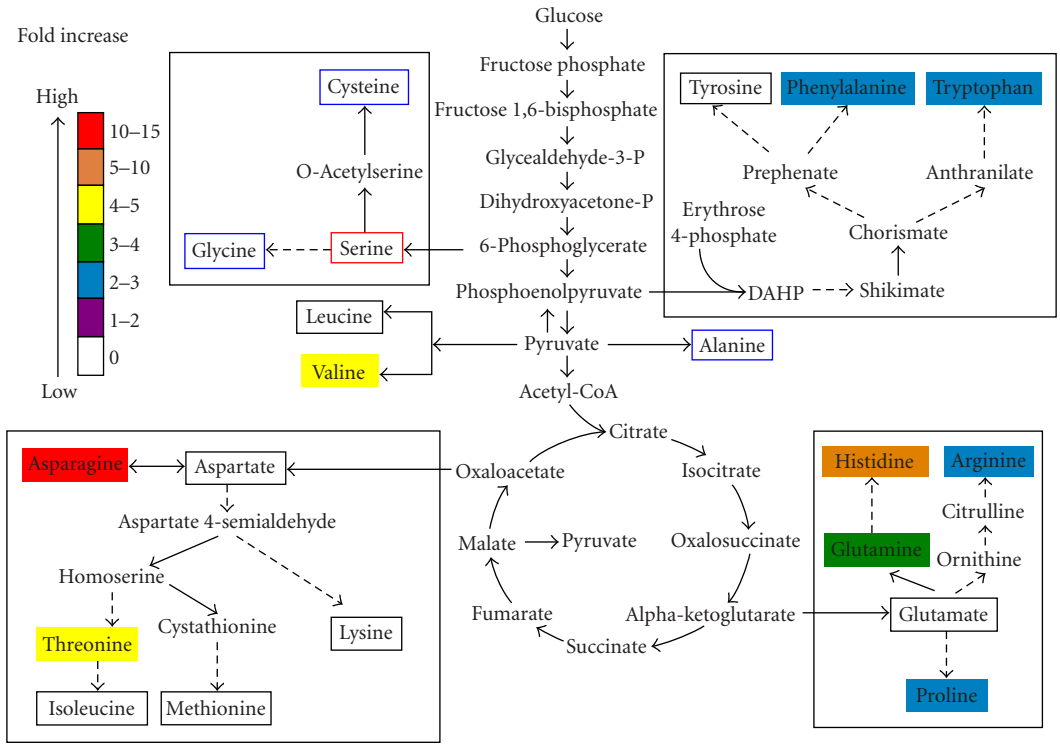
Amino acids, precursors, and derivatives

In leaf extracts, consistent with previous reports of increased free amino acids [12], we found 6 amino acids that increased in abundance (1.3 to 4.5 fold, Figure 4a) in GDH plants. Arginine, phenylalanine, tryptophan, asparagine, glutamine, and histidine were inferred to be altered in abundance. Most of the known pathway intermediates involved in the biosynthesis of protein amino acids were detected, but were not altered in abundance. Four amino acid derivatives changed in abundance (Table 5(a)), one decreased, and three increased. The nonprotein amino acid ornithine increased 2.3 fold.

In roots, 9 amino acids appeared to be increased in abundance in GDH plants by 2 to 11 fold (Table 5(b)). Arginine, phenylalanine, tryptophan, asparagine, glutamine, histidine, proline, threonine, and valine were inferred to be altered in abundance. The root increases in proline, threonine, and valine were not detected in leaves. Many of the known pathway intermediates involved in the biosynthesis of protein amino acids were detected but not altered in abundance, except for the proline precursor delta-pyrroline-5-carboxylate (91, Table 5). No amino



(a)



(b)

FIGURE 4. Metabolites in blue boxes were not detected. Metabolites in red boxes were used as internal standards and therefore detected. Metabolites in black boxes were detected and not changed. (a) Amino acids in leaves increased by *gdhA*. (b) Amino acids in roots increased by *gdhA*. Metabolites that are not protein amino acids are not annotated for changes here.

Table 5. Altered abundance (percentage change) in GDH plants compared to non-GDH plants in (a) amino acid derivatives in leaf extracts, (b) amino acid derivatives in root extracts, (c) sugars and derivatives in leaf extracts, (d) sugars and derivatives in root extracts, (e) fatty acids in leaf extracts, (f) fatty acid derivatives and conjugates in leaf extracts, (g) fatty acids in root extracts, (h) fatty acid derivatives and conjugates in root extracts, (i) compounds of special nitrogen metabolism in leaf extracts, (j) compounds of special nitrogen metabolism in root extracts, (k) nucleic acids in leaf extracts, (l) nucleic acids in root extracts, (m) TCA cycle intermediates and derivatives in leaf extracts, (n) TCA cycle intermediates and derivatives in root extracts, (o) metabolites involved in stress tolerance in leaf extracts, (p) metabolites involved in stress tolerance in root extracts, (q) miscellaneous metabolites in leaf extracts, (r) miscellaneous metabolites in root extracts- part 1, (s) miscellaneous metabolites in root extracts- part 2. [a]: mass is ±1 ppm, or 0.0002–0.00001 d. [b]: % changes are ±2%. N/A denotes not applicable.

(a)

	Empirical formula	Molecular mass[a]	Percentage change [b]
(1) N-alpha-phenylacetyl-glutamine	C13H16N2O4	264.1110	227
(2) 3-aryl-5-oxoproline ethyl ester	C13H15NO3	233.1052	303
(3) 5-methyl-DL-tryptophan	C12H14N2O2	218.1055	40
(4) N-alpha-BOC-L-tryptophan	C16H20N2O4	304.1423	333

(b)

	Empirical formula	Molecular mass[a]	Percent change[b]
(5) N-acetyl-L-tyrosine	C11H13NO4	223.0845	49
(6) PTH-proline	C12H12N2O3	232.0670	43
(7) (gamma-L-glutamyl)-L-glutamine	C10H17N3O6	275.1117	263
(8) N-Benzoyl-L-tyrosine ethylester	C18H19NO4	314.1201	50
(9) 1-[N-(1-carboxy-3-phenylpropyl)-L-lysyl]-L-proline	C21H31N3O5	405.2264	278

(c)

	Empirical formula	Molecular mass[a]	Percent change[b]
(10) 3-deoxy-D-glycero-D-galacto-2-nonulosonic acid	C9H16O9	268.0794	159
(11) Bis-D-fructose 2′,1:2,1′-dianhydride	C12H20O10	324.1056	208

(d)

	Empirical formula	Molecular mass[a]	Percent change[b]
(12) 1,6-anhydro-beta-D-glucopyranose	C6H10O5	162.0528	263
(13) 2-amino-2-deoxy-D-glucose	C6H13NO5	179.0794	276
(14) Sedoheptulose anhydride	C7H12O6	192.0634	909
(15) 3-Deoxy-D-glycero-D-galacto-2-nonulosonic acid	C9H16O9	268.0794	233
(16) 1,6-anhydro-beta-D-glucopyranose 2,3,4-triacetate	C12H16O8	288.0845	588
(17) Bis-D-fructose 2′,1:2,1′-dianhydride	C12H20O10	324.1056	1250

(e)

Common name	Systematic name	Empirical formula	Molecular mass[a]	Degree of saturation	Percent change[b]
(18) Pentadecanoic acid	*n*-pentadecanoic acid	C15H30O2	242.2246	15:0	23
(19) Palmitoleic acid	Hexadecenoic acid	C16H30O2	254.2246	16:1	12
(20) Palmitic acid	Hexadecanoic acid	C16H32O2	256.2402	16:0	30
(21) Linoleic acid	9,12-octadecanedioic acid	C18H32O2	280.2402	18:2	36
(22) Oleic acid	9-octadecenoic acid	C18H34O2	282.2559	18:1	14
(23) Lignoceric acid	Tetracosanoic acid	C24H48O2	368.3654	24:0	15

(f)

Systematic name	Empirical formula	Molecular mass[a]	change[b]
(24) Ethyl tricosanoate	C25H50O2	382.3811	24
(25) Ethyl tetracosanoate	C26H52O2	396.3967	30

Table 5. Continued.

(g)

Common name	Systematic name	Empirical formula	Degree of saturation	Molecular mass[a]	Percent change[b]
(26) Pelargonic acid	*n*-nonanoic acid	C9H18O2	9:0	158.1380	13
(27) Capric acid	*n*-decanoic acid	C10H20O2	10:0	172.1463	13
(28) Undecanoic acid	*n*-hendecanoic acid	C11H22O2	11:0	186.1620	21
(29) Lauric acid	Dodecanoic acid	C12H24O2	12:0	200.1776	14
(30) N/A	Trans-2-tridecenoic acid	C13H24O2	13:1	212.1776	50
(31) N/A	Tridecanoic acid	C13H26O2	13:0	214.1933	22
(32) Undecanedioic acid	N/A	C11H20O4	11:2	216.1362	14
(33) Pentadecanoic acid	*n*-pentadecanoic acid	C15H30O2	15:0	242.2246	6
(34) Palmitoleic acid	Hexadecenoic acid	C16H30O2	16:1	254.2246	29
(35) Palmitic acid	Hexadecanoic acid	C16H32O2	16:0	256.2402	4
(36) Myristic acid	Tetradecanoic acid	C14H26O4	14:2	258.1831	13
(37) Margaric acid	*n*-heptanoic acid	C17H34O2	17:0	270.2559	19
(38) Oleic acid	9,12-octadecanedioic acid	C18H32O2	18:1	282.2559	32
(39) Stearic acid	Octadecenoic acid	C18H34O2	18:0	284.2715	11
(40) N/A	*n*-nonanoic acid	C19H38O2	19:0	298.2872	10
(41) DL-12-hydroxystearic acid	N/A	C18H36O3	18:0	300.2664	196
(42) Tricosanois acid	*n*-tricosanoic acid	C23H46O2	23:0	354.3498	13
(43) Lignoceric acid	Tetracosanoic acid	C24H48O2	24:0	368.3654	5

(h)

Systematic name	Empirical formula	Molecular mass[a]	Percent change[b]
(44) Tetradecanoic acid, 7-oxo-, methyl ester	C15H28O	224.2140	43
(45) (9Z)-(13S)-12,13-epoxyoctadeca-9,11-dienoate	C18H30O3	294.2195	192
(46) 9-Octadecenoic acid, methyl ester	C19H36O2	296.2715	23
(47) Ethyl linoleate	C20H36O2	308.2715	31
(48) (9Z,11E,14Z)-(13S)-hydroperoxyoctadeca-(9,11,14)-trienoate	C18H30O4	310.2144	238
(49) Methyl 12-oxo-trans-10-octadecenoate	C19H34O3	310.2508	25
(50) Octadecanoic acid, ethenyl ester	C20H38O2	310.2872	17
(51) (9Z,11E)-(13S)-13-hydroperoxyoctadeca-9,11-dienoate	C18H32O4	312.2301	194
(52) Octadecanoic acid, 12-oxo-, methyl ester	C19H36O3	312.2664	14
(53) Diethyl tetradecanedioate	C18H34O4	314.2457	19
(54) propyl stearate	C21H32O2	326.3185	18
(55) 5(S)-hydroperoxy-arachidonate	C20H32O4	336.2301	714
(56) Octadecanoic acid, 9,10-epoxy-, allyl ester	C21H38O3	338.2821	10
(57) Ethyl tricosanoate	C25H50O2	382.3811	7
(58) Ethyl tetracosanoate	C26H52O2	396.3967	8
(59) 4,4′-Dimethylcholestatrienol	C29H46O	410.3549	16

(i)

Class amines	Empirical formula	Molecular mass [a]	Percent change[b]
(60) N-caffeoylputrescine	C13H18N2O3	250.1317	196
	Alkaloids		
(61) 8-acetyl quinoline	C11H0NO2	187.0633	227
(62) Scopoletin	C10H8O4	192.0423	244
	Phenolics		
(63) Acetophenone	C8H8O	120.0575	238
(64) 4-hydroxycoumarin	C9H6O3	162.0317	270
(65) N,N-dimethyl-5-methoxytryptamine	C13H18N2O	218.1419	294

Table 5. Continued.

(j)

Class amines	Empirical formula	Molecular mass[a]	Percent change[b]
(66) Epinine	C9H13NO2	167.0946	222
(67) N-caffeoylputrescine	C13H18N2O3	250.1317	19
	Alkaloids		
(68) Coumarin	C9H6O2	146.0368	10
(69) Indole-5,6-quinone	C8H5NO2	147.0393	40
(70) 2-methyl cinnamic acid	C10H202	162.0681	59
(71) 3-acetylaminoquinoline	C11H10N2O	186.0793	34
(72) 7-ethoxy-4-methylcoumarin	C12H12O3	204.0786	36
(73) 4,6-dimethyl-8-tert-butylcoumarin	C15H18O2	230.1307	27
(74) 1-O-hexyl-2,3,5-trimethylhydroquinone	C15H24O2	236.1776	179
	Phenolics		
(75) Acetophenone	C8H8O	120.0575	54
(76) Alpha-hydroxyacetophenone	C8H8O2	136.0524	49
(77) Nicotine	C10H14N2	162.1157	270
(78) Swainsonine	C8H15N2	173.1052	500
(79) (S)-6-hydroxynicotine	C10H14N2O	178.1106	263
	Isoprenoid		
(80) Nopinone	C9H14O	138.1045	20

(k)

	Empirical formula	Molecular mass[a]	Percent change[b]
(81) 2,3-cyclopentenopyridine	C8H9N	119.0735	278
(82) Dihydro-thymine	C6H5N2O2	128.0586	227

(l)

	Empirical formula	Molecular mass[a]	Percent change[b]
(84) Dihydro-thymine	C6H5N2O2	128.0586	238
(85) Uridine	C9H12N2O6	244.0695	400

(m)

	Empirical formula	Molecular mass[a]	Percent change[b]
(86) Fumaric acid, monoethyl ester	C6H8O4	144.0423	56

(n)

	Formula	Mass[a]	Change[b]
(87) Fumaric acid	C4H404	116.0110	270
(88) DL-malic acid	C4H6O5	134.0215	270
(89) Citric acid	C6H8O7	192.0270	385
(90) Fumaric acid monoethyl ester	C6H8O4	144.0423	345

Table 5. Continued.

(o)

	Empirical formula	Molecular mass[a]	Percent change[b]
(91) 3-hydroxy-1-pyrroline-delta-carboxylate	C5H7NO3	129.0426	133

(p)

	Empirical formula	Molecular mass[a]	Percent change[b]
(92) Delta1-pyrroline 2-carboxylate	C5H7NO2	113.0477	217
(93) 3-hydroxy-1-pyrroline-gamma-carboxylate	C5H7NO3	129.0426	244

(q)

(94) N-nitrosopyrrolidine	C4H8N2O	100.0637	152
(95) 2-furylglyoxylonitrile	C6H3NO2	121.0164	182
(96) L-threonate	C4H8O5	136.0372	370
(97) 4-phenyl-2-thiazoleethanamide	C11H12N2S	204.0721	47
(98) Diethyl 1,4 piperazine dicarboxylate	C10H18N2O4	230.1267	54
(99) Hopantenic acid	C10H18NO5	233.1263	34
(100) Menthyl acetoacetate	C14H24O3	240.1725	23
(101) N-methyl-5-allyl-cyclopentylbarbituric acid	C13H16N2O3	248.1161	208
(102) 1-(3-benzoyloxyphenyl)-3-methyl-3-methoxyurea	C16H16N2O4	300.1110	192
(103) 1-(3-benzyloxylphenyl)-3-methyl-3-methoxylurea	C16H20N2O4	304.1350	333
(104) 1,4-Bis((2-((2-hydroxyethyl)amino)ethyl)amino)-9,10-anthracenedione diacetate	C26H32N4O6	496.2322	345

(r)

	Formula	Mass[a]	Change[b]
(105) N-nitrosopyrrolidine	C4H8N2O	100.0637	714
(106) R-4-hydroxy-2-pyrrolidone	C4H7NO2	101.0477	435
(107) 3-methoxy-1,2-propanediol	C4H10O3	106.0630	40
(108) cis-2-hexenoic acid amide	C6H11NO	113.0841	26
(109) 7-oxabicyclo[2.2.1]hept-5-ene-2,3-dione	C6H4O3	124.0160	41
(110) 2-methoxy-3-methyl-pyrazine	C6H8N2O	124.0637	51
(111) Phthalic anhydride	C8H4O3	148.0160	24
(112) Gamma-nonanolactone	C9H16O2	156.1150	43
(113) 1,5-diazatricyclo [4.2.2.2(2,5)]dodecane	C10H18N2	166.0994	625
(114) 2-decenoic acid	C10H18O2	170.1307	56
(115) 2,2,6,6-tetramethyl-N-nitrosopiperidine	C9H18N2O	170.1419	29
(116) 1-acetyl-4-piperidinecarboxylic acid	C8H13NO3	171.0895	270
(117) Decanamide	C10H21NO	171.1623	435
(118) Sulfuric acid dipropyl ester	C6H14N2O8	182.0613	56
(119) o,o′-iminostilbene	C4H11N	193.0892	13
(120) Cyclohexanepropionic acid, 4-oxo-, ethyl ester	C11H18O3	198.1256	25
(121) Cyclooctyl-1,1-dimethylurea	C11H22N2O	198.1732	24
(122) Sebacic acid	C10H18O4	202.1205	16
(123) cis-2,6-di-tert-butylcyclohexanone	C14H26O	210.1984	35
(124) 6-[2-(5-nitrofuranyl)ethenyl]-2-pyridinemethanol	C12H10N2O4	224.0797	213
(125) 5-allyl-5-butylbarbituric acid	C11H16N2O3	224.1161	22

Table 5. Continued.

(s)

	Empirical formula	Molecular mass[a]	Percent change[b]
(126) Isothiocyanic acid 1,4-cyclohexylene-dimethylene ester	C15H24O2	226.0598	31
(127) Tetradecanamide	C14H29NO	227.2249	23
(128) Cedrol methyl ether	C16H28O	236.2140	21
(129) Cyclohexadecanone	C16H30O	238.2297	18
(130) 1,3-di-o-tolylguanidine	C15H17N3	239.1422	400
(131) Menthyl acetoacetate	C14H24O3	240.1725	13
(132) Methocarbamol	C11H15NO3	241.0950	244
(133) N-[2,6-bis(isopropyl)phenyl]-2-imidazolidineimine	C15H23N3	245.1892	345
(134) (-)-ptilocaulin	C15H25N3	247.2048	294
(135) 1-Lauryl-2-pyrrolidone	C16H31NO	253.2406	29
(136) Hexadecanamide	C16H33NO	255.2562	12
(137) Dodecylmalonic acid	C15H28O4	272.1988	46
(138) 4-amino-N-(6-methoxy-4-pyrimidyl)-benzenesulfonamide	C11H12N4O3S	280.0630	20
(139) Rocastine	C13H19N3OS	281.1198	276
(140) Palmoxiric acid	C17H32O3	284.2351	35
(141) Propionic acid, 3-dodecyloxy-2-ethoxy-, methyl ester	C18H36O4	316.2614	556
(142) Benzenesulfonic acid dodecylester	C18H30O3S	326.1916	63
(143) Di(2-ethylhexyl) itaconate	C21H38O4	354.2770	40
(144) 2,2′-ethyledene bis (4,6-di-t-butyl)	C30H45O2	438.3498	12

acid was decreased in abundance in GDH plants but three of the five amino acid derivatives that were changed decreased (Table 5(b)).

Sugars and derivatives

In leaves, consistent with a previous report of a modest increase in free carbohydrates [12], two sugar derivatives appeared to be increased 1.5–2 fold (10,11, Table 5) in GDH plants. In roots (Table 5(d)) six sugars appeared to be increased by 2.3–12.5 fold between GDH and non-GDH plants and included key intermediates involved in the regeneration of ribulose-5-phosphate in the Benson-Calvin cycle. One sugar derivative increased in both leaves and roots (10,15, Table 5).

Fatty acids

In leaves the six fatty acids and two derivatives that appeared to be changed in abundance all decreased in the GDH plants (18–25, Table 5). Two 16-carbon fatty acids (19,20, 16:0 and 16:1, Table 5) and two 18-carbon plant membrane fatty acids were reduced (21,22, 18:1, 18:2, Table 5). These changed fatty acids are minor components of both plant cell membranes and chloroplast membranes. However, α-linolenic acid (18:3), the main constituent of both membranes was not altered in abundance. Two rare unsaturated fatty acids (18 and 23; 15:0 and 24:0) were significantly reduced (Table 5(e)). Two fatty acid derivatives also decreased (Table 5(f)).

In roots, seventeen of the eighteen fatty acids and twelve of the sixteen fatty acid derivatives that appeared to be changed in abundance decreased in GDH plants (26–40, 42,43, Table 5). The decreased fatty acids included 5 of those that decreased in leaves (33–35,38,43, Table 5) but the 18:2 was not decreased. The decreased fatty acid derivatives included both those that decreased in leaves. The four fatty acid derivatives that increased included 3 di-enoates or tri-enoates of 18-C fatty acids that may be biosynthetically related (45,48,51,55, Table 5). None of the common diacylglycerol lipids were detected by FT-ICR-MS so whether decreases in fatty acids were reflected by decreases in lipids is not known.

Special nitrogen metabolism

Six metabolites that appeared to be increased in abundance between GDH and non-GDH plants in leaf extracts were amines (1), alkaloids (2), and phenolics (3), three classes of products derived from special nitrogen metabolism (Table 5(i)).

From four classes of special nitrogen metabolites sixteen appeared to be altered in abundance between GDH and non-GDH plants in roots. There were amines (2), alkaloids (7), phenolics (5), and isoprenoids (1) (Table 5(j)). Five increased and nine decreased. Only N-caffeoylputrescine was altered in both organs. However, it increased in leaves but decreased in roots.

Nucleic acids and derivatives

Only two derivatives of nucleic acids appeared to be increased 2–3 fold in leaf extracts between GDH and non-GDH plants (Table 5(k)). Two compounds were increased in roots more than 2 fold, including the common ribonucleotide uridine (Table 5(l)).

TCA cycle intermediates and derivatives

The monoethyl ester of fumaric acid was the sole metabolite identified that appeared to be changed by GDH in leaves (Table 5(m)). In roots, all four metabolites that changed increased in abundance (2.7–4 fold) including three TCA cycle intermediates, fumaric, malic, and citric acids (Table 5(n)).

Stress-related compounds

Only one member of this group appeared to be altered in leaf extracts (Table 5(o)); two increased more than 2 fold in roots (Table 5(p)).

Miscellaneous

Ten metabolites appeared to be altered in leaves and eight contained nitrogen. Five of the compounds identified in leaf extracts represent known drugs and cigarette toxins (Table 5(q)).

Forty metabolites appeared to be altered in roots and twenty-two contained nitrogen. Among these are five drugs, five flavoring agents, four pesticides, three carcinogens, and five toxins. There were two compounds that were also coordinately altered in leaves: N-nitrosopyrollidone and menthyl acetoacetate.

DISCUSSION

Metabolite analysis

This study used metabolite analysis with FT-ICR-MS [5, 8] to associate phenotype with biochemical changes resulting from endogenous effects of ectopic glutamate synthesis in transgenic plants. The GDH plants were a suitable test for FT-ICR-MS because they have cell composition alterations that result from a specific biochemical alteration in a well-characterized pathway targeting the cellular glutamate pools [12].

The identification of ions and the inference of a metabolite were relatively inefficient, with less than half the ions detected having known metabolites of corresponding masses. The rest of the ions may represent reactions occurring before sample quenching, multiple ionization effects, ion suppression effects, or ion fragmentation [9]. Some of these ions may represent new metabolites not previously reported in plants. An estimate of the extent of artifact ions compared to new products will be a future goal.

Although not reliable or fully quantitative, the changed abundances inferred from ions detected by FT-ICR-MS that appear to correspond to metabolites such as amino acids, sugars, and fatty acids largely agreed with quantitative spectrophotometer assays [12, 17, 18, 19] HPLC separation of sixteen individual amino acids from the methanol soluble, low molecular weight fraction of cell extracts showed eight were significantly changed. Four of the eight amino acids had been inferred to increase in abundance by FT-ICR-MS; the remainder had not been detected as quality peaks. Three amino acids, histidine, valine, and threonine, were not increased as expected from FT-ICR-MS. The difference appears to be due to interference by other ions [9].

Given the only partial agreement among three different measurements of the amino acids, we conclude that the abundance of specific ions in total infusion mass spectra were significantly affected by combined ion suppression effects of all other components, pH and salinity of the solution, flow rate, tip opening, and electrospray current [9]. Further the samples may have differed in the rapidity of turnover of intracellular metabolites, the rate at which metabolism was quenched and the time for which metabolites were separated from the cell debris [10]. The evidence of reproducibility for some amino acid measurements may be related to handling samples simultaneously [9, 10]. Samples analyzed separately either temporally or spatially will be more difficult to compare.

However, the exact masses alone are not sufficient to identify specific compounds unequivocally. Several compounds were identified that are not metabolites in plants (eg, alpha-tert.butoxycarbonyl-L-tryptophan compound no 4, Table 5(a), a synthetic intermediate in peptide synthesis); some artificial pesticide-like metabolites (119,126,127,133,138, Table 5); and metabolites found in insects not plants (eg, no 114, a component of bee royal jelly). Therefore, data from FT-ICR-MS analysis should be used as preliminary evidence to suggest further experiments [8]. In this publication we focused on amino acid metabolism and the effect on central metabolism.

Among the effects detected, those altering amino acid metabolism and fatty acid metabolism were most profound and appear to underlie a doubling of free amino acids and halving of free fatty acid content [12, 13, 18]. In comparison the effects of GDH on carbohydrate metabolism were comparatively trivial and may not solely underlie the increased content reported [12]. The increases in three abundant organic acids may have contributed to the carbohydrate content reported by spectrophotometer assays of reducing sugar content. In addition some of the unidentified ions may have been sugars or carbohydrates.

The majority of ions detected by FT-ICR-MS could not be identified from their predicted formulas or mass. In comparison about 30% of ions identified in plants by GC-MS could not be identified [4, 7, 11]. The unidentified ions detected may represent novel constituents of tobacco leaves or roots [38] or ionization artifacts of MS [5, 6]. Different abundances could also be experimental artifacts. To reduce artifacts we used pooled samples for each genotype from plants grown in an RCB in a growth

chamber; accepted only those ions derived from two ionization methods of the four applied; and by repeating the entire experiment.

Masses of the unidentified metabolite ions are available online at http://www.siu.edu/~pbgc/metaboliteprofiles/GDH/Ntabacum/IONS1-4.html but not discussed further here for brevity.

The concurrence between FT-ICR-MS and spectrophotometer data [12] appears to validate the use of the method for metabolite analyses. However, the informational content of FT-ICR-MS is orders of magnitude greater than other high-throughput methods [1]. FT-ICR-MS detected 2 012 ions that were consistent across ionization methods from each replicated extract. In comparison, tandem MS required several independent extractions to identify 326 metabolites [4] or 88 metabolites [11]. However, there is no doubt that GC-MS in a tandem format is a superior technique for unequivocal identification of ions and therefore metabolites [1]. In addition GC-MS is superior in that ion concentrations can be derived. Both methods suffer from tuning artifacts among spatially separated runs that can only be partly compensated for by internal standards. We conclude that FT-ICR-MS will have a role in functional genomics where sample throughput is more important than chemical identification and relative quantifications, a situation analogous to the decision to employ microarray or macroarray for analysis of the transcriptome.

Amino acid metabolism

Ammonium assimilation fluxes in roots showed that the introduced GDH contributed to total labeling of glutamate regardless of GS inhibition, suggesting that the enzymes compete for NH_4^+. In leaves, the GDH reduced net $^{13}NH_3$ assimilation, possibly by suppression of GS activity [20, 21]. This is consistent with the increase in leaf NH_4^+ reported [12]. However, in the presence of MSX, GDH partially substituted GS by increasing glutamate labeling (7% label incorporated compared to 2.9% for controls). It is possible that the higher Km of GDH for NH_4^+ [22] and the 7–10 fold greater fluxes in nitrogen (resulting from photorespiration) [23] drive the NAD(P)H-dependent GDH reaction forward to produce glutamate in large quantities during GS inhibition [20]. From labeling we conclude that the modifications in transgenic plants are not the product of greatly increased efficiency of nitrogen assimilation by GDH plants. Instead, the glutamate generated in the cytoplasm may result in altered metabolic fluxes and profiles.

Metabolite analysis apparently contradicts ^{13}N flux labeling because the steady-state of extractable glutamate is not altered between GDH and control transgenic roots or leaves. Short-term flux does not always predict steady-state concentration because plants have mechanisms for sensing nitrogen fluxes and maintaining homeostasis [24, 25]. Flux away from glutamate appears to equal the extra flux into glutamate as many major nitrogen sinks were increased and few decreased (in leaves 19 increased and 4 decreased, in roots 29 increased and 17 decreased). Since plant mRNA abundances did not change (data not shown), allosteric effectors of many enzymes may be involved [26]. The effects of GDH expression on phenotype may result from the signaling effects of increased cytosolic glutamate seen in plants grown at low light intensities, *nia* and *rbc* mutants [27] and the status of certain inorganic N compounds [28, 29, 30]. Metabolites derived from nitrate, such as ammonium, glutamine, and glutamate all may act as signals to report on organic N status [25]. Cytosolic glutamate may act directly as a ligand to activate ion channels.

Metabolites that shared C skeletons were coordinately altered in abundance in both roots and leaves in response to GDH activity. Among the amino acids, 4 that derived from alpha-ketoglutarate were coordinately changed in roots and 3 in leaves. Ornithine, the nonprotein amino acid derived from alpha-ketoglutarate, was also increased in roots. Also changed to the same extent in both organs were phosphoenol pyruvate derivatives, phenylalanine, and tryptophan. However, tyrosine, the only other amino acid originating from phosphoenolpyruvate C skeletons, was not altered in abundance. Asparagine was the only amino acid, derived from oxaloacetate, changed in both leaves and roots but threonine increased 4–5 fold in roots. Among amino acids derived from pyruvate C skeletons, valine was increased 4–5 fold in roots but no changes in leucine were seen and alanine was used as an internal standard and therefore changes could not be detected. Similarly, the amino acids derived from 6-phosphoglycerate could not be detected. The pattern of amino acid changes is similar to that in maize endosperm with the opaque mutation [29] where endogenous GDH activity is increased, but different from that caused by photosynthesis [31, 32]. Therefore, the metabolic alterations are GDH specific, not systemic, implying that the effect of GDH on metabolism depends on the metabolism occurring in the cell.

Changes related to water deficit tolerance

Increases in sugar concentrations could also significantly increase the water deficit tolerance [33]. However, the sugars increased by GDH were complex sugars, not the monosaccharides or disaccharides normally associated with tolerance. The notion of sugar sensing is also gaining momentum [25]. The FT-ICR-MS assays would not detect polysaccharides over 700 d, so again flux and steady state may differ.

None of the following compatible solutes were changed in abundance in either leaves or roots [33]: trigonelline, trehalose, dimethylsulfoniopropionate, glycerol, sorbitol, mannitol, choline-O-sulphate, beta alanine betaine, glycinebetaine, prolinebetaine, N-methyl-proline, hydroxyproline, hydroxyprolinebetaine, and pipecolic acid. However, since the association of water deficit tolerance with any single solute is imperfect,

we expect the phenotype was derived from a combination of increased compatible solutes. One or a few of the unidentified metabolites may also participate. Stomatal behavior, GS activity and resistance to photooxidation may contribute to the tolerant phenotype [34]. Plant morphology does not appear to contribute as the root to shoot ratios were not changed [13, 14]. The mechanism by which water deficit tolerance is afforded remains to be unraveled.

Fatty acids

Oil and protein contents are inversely related to each another, to carbohydrate content and to yield in many crop plants. The synthetic pathways for fatty acid, protein, and carbohydrate compete for carbon skeletons [35]. Therefore, the increase in protein and sugar caused by GDH was expected to cause the reduction in fatty acid content observed in leaves and more pronouncedly in roots. The 16-carbon and 18-carbon fatty acids changed were common constituents of the diacylglycerols in plant cell membranes and chloroplast membranes. However, the most abundant fatty acid in cell membranes, α-linolenic acid (18:3), was not altered in abundance in either leaves or roots. The other fatty acids unaltered in abundance are of the 16:3 class. These are mainly found in chloroplast membranes, albeit in quantities far lower than α-linolenic acid (18:3). Interestingly, only the fatty acids whose contribution to plant membrane composition is minor were reduced. The cells of GDH transgenics appear to be regulating closely the abundance of the most common fatty acids necessary for normal cellular function.

The TCA cycle intermediates that are increased by GDH, fumarate, malate, and citrate (Figure 1) may be associated with the redirection of C away from fatty acid synthesis and toward amino acid synthesis. Fumarate and malate are immediate precursors to pyruvate. Citrate is produced from the catabolism of acetyl-CoA.

Special nitrogen metabolism

Special nitrogen metabolites (amines, alkaloids, phenolics, and isoprenoids) may represent more than 50% of the compounds in plants in the 100–700 d range [36]. They provide defense against herbivores, microorganisms, or competing plants and color or scent to attract pollinating insects and seed-dispersing or fruit-dispersing animals. Their nitrogen is derived from ammonium assimilation via the amino acids (the carbon skeletons may derive from many diverse pathways) so it was surprising that none were decreased in leaves and only nine were decreased in roots in GDH plants.

The abundance of just 2 amines (of the 48 detected) altered in response to GDH (Figure 4, Table 5, 60 and 67, 66). Amines are products of arginine or ornithine metabolism (that were affected by GDH) so the unchanged amine contents were surprising. The amine N-caffeoylputrescine that was increased in shoots and decreased in roots by GDH (by transport) accumulates during abiotic or biotic stress [37, 38] and will stabilize histones, stabilize biomembranes, inhibit viral replication, and regulate cellular growth [36]. Such changes directed by GDH may be useful for the economic production of plant secondary metabolites.

The alkaloids that were altered by GDH (9 of 34 detected) were mainly coumarin (68, 72, 73, Table 5) and quinone (61, 69, 71, 74, Table 5) derivatives. Alkaloids occur in about 15% of plant taxa, including *N tabacum* [36]. Most derive from amines that are synthesized from amino acids. They accumulate in tissues that are important for survival and reproduction providing chemical defense. Targets include heart, liver, lung, kidney, CNS, and reproductive organs. Toxic alkaloids may have pharmacological uses at nontoxic doses (eg, 62, 68, Table 5) [36]. Scopoletin (62, Table 5) inhibits *Escherichia coli* O157, is antiviral, is anti-inflammatory (5 fold more than aspirin), and is an asthma treatment [36, 39]. Increasing leaf concentrations 2–3 fold with GDH may be a useful approach to finding new uses for the tobacco crop. Coumarin (68, Table 5), a perfumed liver and lung toxicant [40] was decreased 10 fold by GDH in roots, potentially useful for the manipulation of diets based on root crops.

Some (8 of the 186 detected) phenolic compounds were altered by GDH. The production of phenylpropanoids occurs predominantly from the amino acid phenylalanine [41]. Quinones, monoterpenes, and modified side chains derive from other pathways. The 2-fold increase in phenylalanine caused by GDH (Figure 3) may explain why phenolics are the predominant class of special nitrogen metabolites increased in leaves (3) and roots (3) of GDH transgenics. Phenolics provide mechanical support and barriers; insect attractant or repellents; antioxidants used in leather making; and flavor components in wines and herbal teas. Swainsonine (78, Table 5) is an inhibitor of mannosidase II, used as a cancer therapy [42]. The 5-fold increase in abundance could be useful. Nicotine (synthesized from ornithine), an animal stimulant and insect repellent, was increased in roots (77, Table 5) but not in leaves [41]. Nicotine is synthesized in the roots and transported to the leaves so increased synthesis may not produce a desirable outcome. Nopinone (80, Table 5) was the only isoprenoid affected by GDH but it has no important pharmacological properties [41].

Nucleic acids

The synthesis of nucleic acids from glutamine is a major nitrogen sink in plant cells [36]. GDH did not alter the abundance of the common phosphonucleotides. Uridine was increased 4 fold. Uridine is a precursor of important biosynthetic compounds UMP, UDP, UTP and their glycosyl derivatives. However, these compounds were not altered in abundance. Clearly the altered amino acid fluxes caused by GDH are being directed toward specific pathways and intermediates, leaving others unaffected.

Miscellaneous compounds

The 46 compounds we termed miscellaneous that were altered by GDH in tobacco included 29 that contain nitrogen but structurally cannot be classified with the special nitrogen metabolites [36]. The predominance of N-containing compounds suggests these alterations are directed by GDH-induced metabolism (Table 5(q), Table 5(r), and Table 5(s)). This group of compounds includes some of medicinal relevance (http://www.cieer.org/geirs/); an antihelmitic (98, Table 5); a tumorstatic that binds to nucleic acids (104, Table 5); a vitamin C metabolite that causes increased absorption, cellular uptake, accumulation and reduced excretion (96, Table 5); a nootropic (a drug that enhances mental function; 99, Table 5); treatments for diabetes, high blood pressure and arteriosclerosis (also found in bee royal jelly; 114, Table 5). An inhibitor of neutral sphingomyelinase (117, Table 5); a mycotoxin and an antitumor agent (134, Table 5); and a GABA uptake inhibitor (116, Table 5). Some constituents of cosmetics (135, 113, Table 5), flavoring agents (131, 124, Table 5), and a solvent (110, Table 5) were altered in GDH plants. Altering abundance of these compounds may provide alternate uses for the tobacco crop.

Some pesticide-like metabolites were altered by GDH although the plants were not exposed to pesticides (119,126,127,133,138, Table 5). These compounds may be enzyme substrates occurring naturally in plants that are structurally similar to pesticides. Cataloging metabolites may lead to new leads for pesticidal chemical discovery [1].

Carcinogens and poisons were primarily altered in abundance in the roots (100 and 131, 105, 107, 108, 111, 115, 128, 132, Table 5) consistent with the root synthesis of these compounds early in development and later translocation to the leaf [41]. The carcinogens and cigarette components detected are specific to tobacco and most probably part of its inherent secondary metabolism. Detection of carcinogens and poisons may serve to validate the use of FT-ICR-MS and suggests applications in the association of smoking with cancer incidence.

Plant pigments, haem, and other porphyrins are major sinks for glutamate source pools in plants [27]. However, the glutamate flux perturbation caused by GDH does not alter the regulation or intermediates of pigment biosynthesis.

We conclude that GDH can be useful for plant metabolic engineering to increase or decrease the yield of a large number of chemical compounds. GDH may be a useful tool as the pharmaceutical industry discovers new plant-derived compounds of therapeutic value.

The work presented here demonstrated that metabolite analyses by FT-ICR-MS provide a useful tool for the analysis of cryptic phenotypes in transgenic plants. The analysis of data from extracts without derivatization allows analysis of the relationships between various metabolites and the equivalence of samples. If there are 40 000 different molecules among all extant plant species in the range of 100–700 d [36], cataloging them by means other than FT-ICR-MS would be a mammoth task [3, 4]. Assuming there are 3–4 thousand different molecules in individual plant species in the range of 100–700 d [4, 11] we will have sampled about 50%–60% (2 012) in two analyses (replicated). However, it is clear from our data that about 50% of the molecules detected are not in the databases we interrogated. Therefore, estimates of the chemical diversity of plant may be grossly underestimated. The development of a cell map and exploration of metabolic instantiations with that map will be impossible without cataloging the consequences of metabolism accurately.

The sensitivity and resolution of FT-ICR-MS provides a useful method for cataloging chemical diversity. Within the existing limits, differences may be measured between samples comprising more than ten thousand cells. Therefore, the occurrence of novel compounds of biomedical significance in individuals, populations, species, and genera may be cataloged.

MATERIALS AND METHODS

Gene manipulations and construction of plasmids

To examine the effects of NADPH-GDH in plants we used three lines, GDH10, GUS, and BAR, described previously [12, 13, 14, 15, 17]. GDH10 is a well-characterized independent line of *Nicotiana tabacum* var "Petite Havana SR1" that expresses the *E coli gdhA* gene. The line represents an early regenerant and lacks noticeable variation from the wild type under normal growth conditions. The *gdhA* gene inserted in GDH10 plants has an architecture and segregation pattern consistent with a single site of insertion. The gene is under the control of the CaMV 35S promoter. Transcript abundances are equal when comparing roots and leaves. Enzyme activity is found in the cytoplasm but not plastids and is equal in roots and leaves. GUS is a well-characterized independent line of *N tabacum* var Petite Havana SR1 that expresses the modified *gusA* gene. The line represents an early regenerant and lacks noticeable variation from the wild type under normal growth conditions. The *gusA* gene inserted in GUS plants has an architecture and segregation pattern consistent with a single site of insertion. The gene is under the control of the CaMV 35S promoter. Enzyme activity is found in the cytoplasm but not plastids and is equal in roots and leaves. BAR is a well-characterized independent line of *N tabacum* var Petite Havana SR1 that expresses the *S hygroscopicus bar* gene. The line represents an early regenerant and lacks noticeable variation from the wild type under normal growth conditions. The *bar* gene inserted in BAR plants has an architecture and segregation pattern consistent with a single site of insertion. The gene is under the control of the CaMV 35S promoter. Enzyme activity provides tolerance to phosphinothricin herbicide to roots, leaves, and cell culture derived from them. BAR

and GUS were chosen as adequate controls because they were not significantly different from wild-type SR1 across a wide range of growth conditions, locations, and years [12, 13, 14, 15, 17].

Seeds of the lines described and clones used for transformation are freely available on request and are being widely used for transformation of other plant species.

Plant material and growth conditions

Tobacco seeds were obtained from the seed stocks at the Agriculture Research Center, Southern Illinois University at Carbondale (Carbondale, Ill). Seeds were sown in 4-inch pots [14] containing a mixture of sand and soil (1:1). Seedlings were thinned to one plant per pot, watered daily, and grown on unshaded benches and in the Horticulture Research Center, Southern Illinois University, from 9/99 to 9/03. The conditions for the growth of plants for ^{13}N labeling, and metabolomics are described in the coming sections. Seeds of each line used are available on request.

Preparation of cell free extracts and GDH assays

GDH assays were performed exactly as described [12]. All preparative steps were carried out at 4°C. The specific activity of aminating NADPH-GDH was quantified by measuring the rate of oxidation of NADPH dependent on reductive amination of alpha-ketoglutarate. Assays were performed at 25°C. The amount of protein in the extracts was determined by Bradford assay.

Labeling of the glutamate pool by ^{13}N

Three individual plants were fed $^{13}NH_4^+$ for 15 minutes via hydroponic solutions for root feeding and via excised stems for leaf feeding, then treated with liquid N, ground up, extracted with distiled water, filtered through glass wool, and separated on an anion-exchange column (Dowex 2X8-100) which retained glutamate. The eluate was washed through the column with another 10 mL of distiled water and passed through a cation-exchange resin (Dowex 50WX8-100) which bound NH_4^+. Glutamine came through in the eluate. The columns were washed with 10 mL of 2M KCl to elute glutamate (anion-exchange column) and NH_4^+ (cation-exchange column). Eluates were collected in 20 mL scintillation vials and counted in a Canberra Packard gamma counter that was automatically corrected for decay time (^{13}N has a half-life of 10 minutes). The percent label incorporated was calculated using the following formula: [{(percent ^{13}N as glutamate in the presence of MSX by GDH10 line) − (percent ^{13}N as glutamate in the presence of MSX by non-GDH line)}/(percent ^{13}N as glutamate in the absence of MSX by the GDH10 line)].

Preparation of metabolite extracts for FT-ICR-MS assays

Three pooled leaf and root samples from each control and transgenic genotype were used to remove spatial and genetic variation not associated with GDH activity. About 100 mg of tissue was ground to which 1.0 mL of 50/50 (v/v) methanol/0.1% (w/v) formic acid was added [8]. The samples were homogenized and centrifuged. The supernatant was used for the analyses. Each sample was mixed with a known and equal amount of a standard mix of serine, tetra-alanine, reserpine, Hewlett-Packard tuning mix, and the adrenocorticotrophic hormone fragment 4–10. These internal calibration compounds produced 4–5 ions of mass encompassing the range reported allowing for control of spectra used for mass reports. The internal calibration compound peak area was used to detect non-biological variations in abundance reported allowing for control of spectra used for the quantities reported. All analytes we purchased from Sigma-Aldrich (St Louis, Mo) and used without further purification.

FT-ICR-MS assays

Briefly, we used the Bruker Daltonics APEX III FT-ICR-MS equipped with a 7.0 Tesla magnet, electrospray, and APCI ionization sources [8]. Both positive and negative ionizations were carried out. Tips were prepared as previously described [8]. For negative ionization, samples were introduced by capillary, diluted 1:19 in 50% (v/v) methanol, 0.2% (v/v) formic acid, 49.8% (v/v) water. For positive ionization, samples were introduced by capillary, diluted 1:19 in 50% (v/v) methanol, 0.2% (v/v) ammonium hydroxide, 49.8% (v/v) water. Flow rates were 5 μL/min for electrospray and 100 μL/min for APCI ionization sources. ESI, APCI, and ion transfer conditions were optimized using a standard mix of serine, tetra-alanine, reserpine, Hewlett-Packard tuning mix, and the adrenocorticotrophic hormone fragment 4–10. Instrument conditions were optimized for ion intensity and broadband accumulation over the mass range of 100–1000 d. One-megaword data files were acquired and a sinm data transformation was performed prior to Fourier transform and magnitude calculations.

(a) Calibration

All samples were internally calibrated for mass accuracy over the approximate mass range of 100–1000 d using a mixture of the above-mentioned standards. The results for each ionization method can be viewed at http://www.siu.edu/~pbgc/metabolite-profiles/GDH/Ntabacum/ions1-4.html.

All mass deviances from the standard curves were less than 1.0 ppm over the mass range studied, although most of them were typically in the 0.1 to 0.2 ppm range. The value for each peak reported by each ionization method can be viewed at http://www.siu.edu/~pbgc/metabolite-profiles/GDH/Ntabacum/ions1-4.html.

(b) Matrix effects and reproducibility

Mass spectra were recorded by averaging 10 single spectra of 3-seconds acquisition time each. Some

suppression was observed but based on several random samples, spectrum to spectrum fluctuations of signal intensity ratios varied from one another by less than 30%. This value was used to indicate the confidence of each data point. Absolute interference is reported for each ion. It ranged from $1.01E^{+06}$ to $5.82E^{+08}$. The average noise peak was $1.01E^{+06}$. Data is an open source, each value and each spectra can be downloaded from http://www.siu.edu/~pbgc/metabolite-profiles/GDH/.

Preparation of database searching for FT-MS assays

(a) Empirical formula inference

Empirical formulas were inferred for those ions for which the area under the peak changed between treatments. Excluded were peaks with inaccurate mass estimates and peaks with multiple likely empirical formulas. Final empirical formulas and masses for the metabolite inferred from each ion followed the addition or subtraction of a single hydrogen ion or electron depending on the mode of ionization used. An assumption made was that all ions represented single ionization events. When there were specific metabolites that we were interested in evaluating, we used the spreadsheet calculator to identify corresponding ion mass to charge ratios. We then manually examined the raw peak list for the corresponding ion mass to charge ratio.

(b) Database searching

We identified compounds by manually interrogating two publicly available databases, one at Chemfinder (http://chemfinder.cambridgesoft.com) and the second at NIST (http://webbook.nist.gov/chemistry/mwser.html). As the mass of the metabolites increases, the number of possible isomer combinations increases. Determination of the isotope expected in tobacco extracts was made manually with reference to plant biochemistry texts and databases. We did not use Phenomenome PLC. proprietary software (Saskatchewan, Canada), only publicly available databases were used so that data and databases of ions would remain open source.

ACKNOWLEDGMENTS

We thank Dr. Rafiqa Ameziane for valuable discussions and help with experiments. Plant materials were developed with a grant from the Herman Frasch Foundation. Analyses were supported by grants from the Illinois-Missouri Biotechnology Alliance and the Council for Food and Agricultural Research. We thank all at Phenomenome for technical assistance with FT-ICR-MS.

REFERENCES

[1] Glassbrook N, Ryals J. A systematic approach to biochemical profiling. *Curr Opin Plant Biol*. 2001;4(3):186–190.

[2] Katona ZF, Sass P, Molnár-Perl I. Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography mass spectrometry. *J Chromatogr*. A 1999;847(1-2):91–102.

[3] Adams MA, Chen ZL, Landman P, Colmer TD. Simultaneous determination by capillary gas chromatography of organic acids, sugars, and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *Anal Biochem*. 1999;266(1):77–84.

[4] Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics [published correction appears in Nat Biotechnol. 2000;19(2):173]. *Nat Biotechnol*. 2000;18(11):1157–1161.

[5] Angotti M, Maunit B, Muller JF, Bezdetnaya L, Guillemin F. Characterization by matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry of the major photoproducts of temoporfin (m-THPC) and bacteriochlorin (m-THPBC). *J Mass Spectrom*. 2001;36(7):825–831.

[6] Shen Y, Tolic N, Zhao R, et al. High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. *Anal Chem*. 2001;73(13):3011–3021.

[7] Marshall AG, Hendrickson CL, Shi SD. Scaling MS plateaus with high-resolution FT-ICRMS. *Anal Chem*. 2002;74(9):252A–259A.

[8] Aharoni A, Ric de Vos CH, Verhoeven HA, et al. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS*. 2002;6(3):217–234.

[9] Schmidt A, Karas M, Dulcks T. Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI? *J Am Soc Mass Spectrom*. 2003;14(5):492–500.

[10] Allen J, Davey HM, Broadhurst D, et al. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol*. 2003;21(6):692–696.

[11] Roessner U, Willmitzer L, Fernie AR. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol*. 2001;127(3):749–764.

[12] Ameziane R, Bernhard K, Lightfoot DA. Expression of the bacterial *gdhA* gene encoding a NADPH glutamate dehydrogenase in tobacco affects plant growth and development. *Plant and Soil*. 2000;221(1):47–57.

[13] Ameziane R, Bernhardt K, Lightfoot DA. Expression of the bacterial *gdhA* gene encoding a glutamate dehydrogenase in tobacco and corn increased tolerance to the phosphinothricin herbicide. In: Martins-Loucao MA, Lips SH, eds. *Nitrogen in a Sustainable*

Ecosystem: From the Cell to the Plant. Leiden, The Netherlands: Backhuys Publishers; 2000:339–343.
[14] Mungur R. *Metabolic Profiles of GDH transgenic crops* [MS thesis]. Carbondale, Ill: SIUC; 2002:189.
[15] Mungur R, Glass AD, Wood AJ, Lightfoot DA. Increased water deficit tolerance in *Nicotiana tabacum* expressing the *Escherichia coli* glutamate dehydrogenase gene. *Plant Cell Physiology*. In press.
[16] Raamsdonk LM, Teusink B, Broadhurst D, et al. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol*. 2001;19(1):45–50.
[17] Lightfoot DA, Long LM, Vidal ME. Plants containing the *gdhA* gene and methods of use thereof. US patent 5 998 700. 1999.
[18] Lightfoot DA, Long LM, Vidal ME. Plants containing the *gdhA* gene and methods of use thereof. US patent 6 329 573, 2001.
[19] Schmidt RR, Miller P. Polypeptides and polynucleotides relating to alpha and beta subunits of lutamate dehydrogenase and methods of use. US patent 5 879 941. 1999.
[20] Melo-Oliveira R, Oliveira IC, Coruzzi GM. *Arabidopsis* mutant analysis and gene regulation define a nonredundant role for glutamate dehydrogenase in nitrogen assimilation. *Proc Natl Acad Sci U S A*. 1996;93(10):4718–4723.
[21] Becker TW, Carrayol E, Hirel B. Glutamine synthetase and glutamate dehydrogenase isoforms in maize leaves: localization, relative proportion and their role in ammonium assimilation or nitrogen transport. *Planta*. 2000;211(6):800–806.
[22] Wootton JC. Re-assessment of ammonium-ion affinities of NADP-specific glutamate dehydrogenases. Activation of the *Neurospora crassa* enzyme by ammonium and rubidium ions. *Biochem J*. 1983;209(2):527–531.
[23] Keys AJ, Bird IF, Cornelius MJ, Lea PJ, Wallsgrove RM, Miflin BJ. Photorespiratory nitrogen cycle. *Nature*. 1978;275:741–743.
[24] Aubert S, Bligny R, Douce R, Gout E, Ratcliffe RG, Roberts JK. Contribution of glutamate dehydrogenase to mitochondrial glutamate metabolism studied by (13)C and (31)P nuclear magnetic resonance. *J Exp Bot*. 2001;52(354):37–45.
[25] Coruzzi GM, Zhou L. Carbon and nitrogen sensing and signaling in plants: emerging "matrix effects." *Curr Opin Plant Biol*. 2001;4(3):247–253.
[26] Lea PJ, Robinson SA, Stewart GR. The enzymology and metabolism of glutamine, glutamate and asparagine. In: Miflin BJ, Lea PJ, eds. *The Biochemistry of Plants*. New York, NY: Academic Press; 1990:121–159.
[27] Stitt M, Muller C, Matt P, et al. Steps towards an integrated view of nitrogen metabolism. *J Exp Bot*. 2002;53(370):959–970.
[28] Wang R, Guegler K, LaBrie ST, Crawford NM. Genomic analysis of a nutrient response in *Arabidopsis* reveals diverse expression patterns and novel metabolic and potential regulatory genes induced by nitrate. *Plant Cell*. 2000;12(8):1491–1509.
[29] Wang X, Larkins BA. Genetic analysis of amino acid accumulation in opaque-2 maize endosperm. *Plant Physiol*. 2001;125(4):1766–1777.
[30] Glass ADM, Britto DT, Kaiser BN, et al. The regulation of nitrate and ammonium transport systems in plants. *J Exp Bot*. 2002;53(370):855–864.
[31] Geiger M , Walch-Liu P, Engels C, et al. Enhanced carbon dioxide leads to a modified diurnal rhythm of nitrate reductase activity in older plants, and a large stimulation of nitrate reductase activity and higher levels of amino acids in young tobacco plants. *Plant Cell Environ*. 1998;21(3):253–268.
[32] Noctor G, Novitskaya L, Lea PJ, Foyer CH. Co-ordination of leaf minor amino acid contents in crop species: significance and interpretation. *J Exp Bot*. 2002;53(370):939–945.
[33] Chen TH, Murata N. Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. *Curr Opin Plant Biol*. 2002;5(3):250–257.
[34] Hoshida H, Tanaka Y, Hibino T, et al. Enhanced tolerance to salt stress in transgenic rice that overexpresses chloroplast glutamine synthetase. *Plant Mol Biol*. 2000;43(1):103–111.
[35] Fell DA, Wagner A. The small world of metabolism. *Nat Biotechnol*. 2000;18(11):1121–1122.
[36] Wink M. Special nitrogen metabolism. In: Dey PM, Harborne JB, eds. *Plant Biochemistry*. London, UK: Academic Press; 1997:439–486.
[37] Balint R, Cooper G, Staebell M, Filner P. N-caffeoyl-4-amino-n-butyric acid, a new flower-specific metabolite in cultured tobacco cells and tobacco plants. *J Biol Chem*. 1987;262(23):11026–11031.
[38] Baumert A, Mock HP, Schmidt J, Herbers K, Sonnewald U, Strack D. Patterns of phenylpropanoids in non-inoculated and potato virus Y-inoculated leaves of transgenic tobacco plants expressing yeast-derived invertase. *Phytochemistry*. 2001;56(6):535–541.
[39] Duncan SH, Flint HJ, Stewart CS. Inhibitory activity of gut bacteria against *Escherichia coli* O157 mediated by dietary plant metabolites. *FEMS Microbiol Lett*. 1998;164(2):283–288.
[40] Born SL, Caudill D, Fliter KL, Purdon MP. Identification of the cytochromes P450 that catalyze coumarin 3,4-epoxidation and 3-hydroxylation. *Drug Metab Dispos*. 2002;30(5):483–487.
[41] Strack D. Phenolic metabolism. In: Dey PM, Harborne JB , eds. *Plant Biochemistry*. London, UK: Academic Press; 1997:387–416.
[42] Rooprai HK, Kandanearatchi A, Maidment SL, et al. Evaluation of the effects of swainsonine, captopril, tangeretin and nobiletin on the biological behaviour of brain tumour cells in vitro. *Neuropathol Appl Neurobiol*. 2001;27(1):29–39.

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 215–225 • DOI: 10.1155/JBB.2005.215

REVIEW ARTICLE

Finding Groups in Gene Expression Data

David J. Hand and Nicholas A. Heard

Department of Mathematics, Faculty of Physical Sciences, Imperial College, London SW7 2AZ, UK

Received 11 June 2004; revised 24 August 2004; accepted 24 August 2004

The vast potential of the genomic insight offered by microarray technologies has led to their widespread use since they were introduced a decade ago. Application areas include gene function discovery, disease diagnosis, and inferring regulatory networks. Microarray experiments enable large-scale, high-throughput investigations of gene activity and have thus provided the data analyst with a distinctive, high-dimensional field of study. Many questions in this field relate to finding subgroups of data profiles which are very similar. A popular type of exploratory tool for finding subgroups is cluster analysis, and many different flavors of algorithms have been used and indeed tailored for microarray data. Cluster analysis, however, implies a partitioning of the entire data set, and this does not always match the objective. Sometimes pattern discovery or bump hunting tools are more appropriate. This paper reviews these various tools for finding interesting subgroups.

INTRODUCTION

Microarray gene expression studies are now routinely used to measure the transcription levels of an organism's genes at a particular instant of time. These mRNA levels serve as a proxy for either the level of synthesis of proteins encoded by a gene or perhaps its involvement in a metabolic pathway. Differential expression between a control organism and an experimental or diseased organism can thus highlight genes whose function is related to the experimental challenge.

An often cited example is the classification of cancer types (Golub et al [1], Alizadeh et al [2], Bittner et al [3], Nielsen et al [4], Tibshirani et al [5], and Parmigiani et al [6]). Here, conventional diagnostic procedures involve morphological, clinical, and molecular studies of the tissue, which both are highly subjective in their analysis and cause inconvenience and discomfort to the patient. Microarray experiments offer an alternative (or additional), objective means of cell classification through some predetermined functionals of the gene expression levels for a new tissue sample of an unknown type. Whilst potentially very powerful, the statistical robustness of these methods is still hampered by the "large p, small n" problem; a microarray slide can typically hold tens of thousands of gene fragments whose responses here act as the predictor variables (p), whilst the number of patient tissue samples (n) available in such studies is much less (for the above examples, 38 in Golub et al, 96 in Alizadeh et al, 38 in Bittner et al, 41 in Nielsen et al, 63 in Tibshirani et al, and 80 in Parmigiani et al).

More generally, beyond such "supervised" classification problems, there is interest in identifying groups of genes with related expression level patterns over time or across repeated samples, say, even within the same classification label type. Typically one will be looking for coregulated genes showing similar expression levels across the samples, but equally we may be interested in anticorrelated genes showing diametric patterns of regulation (see, eg, Dhillon et al [7]) or even genes related through a path of genes with similar expression (Zhou et al [8]). In the case of classification of cancer, these "unsupervised" studies can give rise to the discovery of new classifications which may be morphologically indistinguishable but pathogenetically quite distinct. In general, they may shed light on unknown gene functions and metabolic pathways.

A common aim, then, is to use the gene expression profiles to identify groups of genes or samples in which the members behave in similar ways. In fact, that task description encompasses several distinct types of objectives.

Firstly, one might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Implicit in this is the assumption that there do exist groups such that members of a given group have similar patterns which are rather different from the patterns exhibited by members of the other groups. The aim,

Correspondence and reprint requests to David J. Hand, Department of Mathematics, Faculty of Physical Sciences, Imperial College, London SW7 2AZ, UK, E-mail: d.j.hand@imperial.ac.uk

then, is to "carve nature at the joints," to identify these groups. Statistical tools for locating such groups go under the generic name of cluster analysis, and there are many such tools.

Secondly, one might simply want to partition the data set to assign genes to groups such that each group contains genes with similar expression profiles, with no notion that the groups are "naturally occurring" or that there exist "joints" at which to carve the data set. This exercise is termed dissection analysis (Kendall [9]). The fact that the same tools are often used for cluster analysis and dissection analysis has sometimes led to confusion.

Thirdly, one might simply want to find local groups of genes which exhibit similar expression profiles, without any aim of partitioning the entire data set. Thus there will be some such local groupings, but many, perhaps most, of the genes will not lie in any of these groups. This sort of exercise has been termed *pattern discovery* (Hand et al [10]).

Fourthly, one might wish to identify groups of genes with high variations over the different samples or perhaps dominated by one label type in a supervised classification setting. Methods for identifying such groups which start with a set of genes and sequentially remove blocks of the genes until some criterion is optimised have been termed by Hastie et al [11] as "gene shaving."

Fifthly, in *pattern matching* one is given a gene a priori, with the aim being to find other genes which have similar expression profiles. Technically, solutions to such problems are similar to those arising in nucleotide sequencing, with more emphasis on imprecise matches.

Sixthly, in supervised classification, there is the case described above where samples of genes are provided which belong to each of several prespecified classes, and the aim is to construct a rule which will allow one to assign new genes to one of these classes purely on the basis of its expression profile (see Golub et al [1]).

Of these objectives, cluster analysis and pattern discovery both seek to say something about the intrinsic structure of the data (in contrast to, eg, dissection and pattern matching) and both are exploratory rather than necessarily being predictive (in contrast to, eg, supervised classification). This means that these problems are fundamentally open ended: it is difficult to say that a tool will never be useful under any circumstances. Perhaps partly because of this, a large number of methods have been developed. In the body of this paper we describe tools which have been developed for cluster analysis and pattern discovery, since these are intrinsically concerned with finding natural groups in the data, and we summarise their properties. We hope that this will be useful for researchers in this area, since two things are apparent: (i) that the use of such methods in this area is growing at a dramatic rate and (ii) that often little thought is given about the appropriateness of the choice of methods. An illustration of the last point is given by the fact that different cluster analysis algorithms are appropriate for detecting different kinds of cluster structure, and yet it is clear that often the choice of methods has been a haphazard one, perhaps based on software availability or programming ease, rather than an informed one.

In the "microarray experiments" section we give an introduction to microarray technology and discuss some of the issues that arise in its analysis. The "cluster analysis" and "pattern discovery" sections detail clustering and pattern discovery methods, respectively; in the case of clustering, examples are given of situations where these techniques have been applied to microarray data, and for pattern discovery we suggest how these methods could carry across to this area. Finally some conclusions are given.

MICROARRAY EXPERIMENTS

There are two main microarray technologies, complementary DNA (cDNA) and oligonucleotide, though both work on the same principle of attaching sequences of DNA to a glass or nylon slide and then hybridising these attached sequences with the corresponding DNA or (more commonly) RNA in a sample of tissue through "complementary binding." The two technologies differ according to the type of "probe" molecules used to represent genes on the array. With cDNA microarrays genes are represented by PCR-amplified (polymerase chain reaction) DNA sequences spotted onto a glass slide; with oligonucleotide arrays, between 16 and 20 complementary subsequences of 25 base pairs from each gene are attached to the chip by photolithography, together known as the perfect match (PM), along with the same sequences altered slightly at the middle bases, known as the mismatch (MM), for factoring out nonspecific binding. As it is difficult to measure the amount of PCR product in the former case, it follows that for cDNA microarrays we can only achieve relative expression levels of two or more samples against one another, whereas for oligonucleotide arrays absolute measurements are taken, such as mean or median of PM-MM or log(PM/MM).

After hybridisation a fluorescence image of the microarray is produced using a scanner, which is usually a laser confocal microscope using excitation light wavelengths to generate emission light from appropriate fluors. This image is pixelated and image analysis techniques are used to measure transcript abundance for each probe and hence give an overall expression score. Finally a normalisation procedure (Dudoit et al [12], Yang et al [13], Irizarry et al [14], and Bolstad et al [15]) is used to remove any systematic variations in the expression scores such as background correction and allow for the effect of location of the probe on the slide (smudges). Besides the difficulties of these procedures, there are many other sources of error such as noncomplementary binding, instability from small expression levels of a gene in both samples, and missing values (Troyanskaya et al [16]).

The resulting low signal-to-noise ratio of microarray experiments means most interest is focused on multiple slide experiments, where each hybridisation process

is performed with tissue samples possibly from the same (replicate data) or different experimental conditions, allowing us to "borrow strength." For life-cycle processes time-course experiments are also popular, where expression levels of an experimental subject are measured at a sequence of time points to build up a temporal profile of gene regulation.

A microarray experiment can measure the expression levels of tens of thousands of genes simultaneously. However, they can be very expensive. Therefore when it comes to data analysis, there is a recurring problem of high dimension in the number of genes and only a small number of cases. This is a characteristic shared by spectroscopic data, which additionally have high correlations between neighbouring frequencies; analogously for microarray data, there is evidence of correlation of expression of genes residing closely to one another on the chromosome (Turkheimer et al [17]). Thus when we come to look at cluster analysis for microarray data, we will see a large emphasis on methods which are computationally suited to cope with the high-dimensional data.

CLUSTER ANALYSIS

The need to group or partition objects seems fundamental to human understanding: once one can identify a class of objects, one can discuss the properties of the class members as a whole, without having to worry about individual differences. As a consequence, there is a vast literature on cluster analysis methods, going back at least as far as the earliest computers. In fact, at one point in the early 1980s new ad hoc clustering algorithms were being developed so rapidly that it was suggested there should be a moratorium on the development of new algorithms while some understanding of the properties of the existing ones was sought. In fact, without this hiatus in development occurring, a general framework for such algorithms was gradually developed.

Another characteristic of the cluster analysis literature, apart from its size, is its diversity. Early work appeared in the statistics literature (where it caused something of a controversy because of its purely descriptive and noninferential nature—was it really statistics?), the early computational literature, and, of course, the biological and medical literature. Biology and medicine are fundamentally concerned with taxonomies and diagnostic and prognostic groupings.

Later, nonheuristic, inferential approaches to clustering founded on probability models would appear. These models fit a mixture probability model to the data to obtain a "classification likelihood," with the similarity of two clusters determined by the change in this likelihood that would be caused by their merger. In fact, most of the heuristic methods can be shown to be equivalent to special cases of these "model-based" approaches. Reviews of model-based clustering procedures can be found in Bock [18], Bensmail et al [19], and Fraley and Raftery [20]. Model-based clustering approaches allow the choice of clustering method and number of clusters to be recast as a statistical model choice problem, and, for example, significance tests can be carried out.

More recently, research in machine learning and data mining has produced new classes of clustering algorithms. These various areas are characterised by their own emphases—they are not merely reinventing the clustering wheel (although, it has to be said, considerable intellectual effort would be saved by more cross-disciplinary reading of the literature). For example, data mining is especially concerned with very large data sets, so that the earlier algorithms could often not be applied, despite advances in computer storage capabilities and processing speed. A fundamental problem is that cluster analysis is based on pairwise similarities between objects and the number of such distances increases as the square of the number of objects in the data set does.

Cluster analysis is basically a data exploration tool, based solely on the underlying notion that the data consist of relatively homogeneous but quite distinct classes. Since cluster analysis is concerned with partitioning the data set, usually each object is assigned to just one cluster. Extensions of the ideas have been made in "soft" clustering, whereby each object may partly belong to more than one cluster. Mixture decomposition, of course, leads naturally to such a situation, and an early description of these ideas (in fact, one of the earliest developments of a special case of the expectation-maximisation (EM) algorithm) was given by Wolfe [21].

Since the aim of cluster analysis is to identify objects which are similar, all such methods depend critically on how "similarity" is defined (note that for model-based clustering this follows automatically from the probability model). In some applications the raw data directly comprise a dissimilarity matrix (eg, in direct subjective preference ratings), but gene expression data come in the form of a gene $\times$ variable data matrix, from which the dissimilarities can be computed. In many applications of cluster analysis, the different variables are not commensurate (eg, income, age, and height when trying to cluster people) so that decisions have to be made about the relative weight to give to the different components. In gene expression data, however, each variable is measured on the same scale. One may, nonetheless, scale the variables (eg, by the standard deviation or some robust alternative) to avoid variables with a greater dispersion playing a dominant role in the distance measure. Note that in the case of model-based clustering methods, however, the reverse is true; different levels of variability for each variable are easy to incorporate into the models whereas the likelihood will be much harder to write down for data rescaled in this way. Likewise, it is worthwhile considering transforming the variables to remove skewness, though, in the case of gene expression data based on the log of a ratio, this may not be necessary or appropriate. Reviews of distance measures are given in Gower [22] and Gordon [23].

Briefly, distance metrics used to define cluster dissimilarity are usually either geometric or correlation based. Variations of the former theme include Euclidean, Manhattan, and Chebychev distances, and of the latter Pearson and Spearman correlations, for example. In the field of gene expression clustering, Eisen et al [24] used an uncentred correlation-based distance measure which takes into account both the "shape" of the gene expression profile and each gene's overall level of expression, and this measure has now been widely adopted.

As noted above, certain types of gene expression clustering problems, such as clustering tissue samples on the basis of gene expression, involve relatively few data points and very large numbers of variables. In such problems especially, though also more generally, one needs to ask whether all the variables contribute to the group structure—and, if not, whether those that do not contribute serve to introduce random variation such that the group structure is concealed (see, eg, Milligan [25], DeSarbo and Mahajan [26], and Fowlkes et al [27]). One can view the problem as that of choosing the distance measure so that the irrelevant variables contribute nothing to the distance, that is, such variables are given a weight of zero if the distance consists of a weighted combination of contributions from the variables. There is a large and growing literature devoted to this problem of selecting the variables. See, for example, De Soete et al [28], De Soete [29, 30], Van Buuren and Heiser [31], and Brusco and Cradit [32]. More generally, of course, one might suspect that different cluster structures occurred in different subsets of variables. This would be the case, for example, if people could suffer from sets of nonmutually exclusive diseases (eg, different pulmonary diseases on the one hand, and psychiatric syndromes on the other). In this case one would ideally like to search over cluster structures and over subsets of variables. In a recent paper, Friedman and Meulman [33] describe a related problem in which, although a single partitioning is sought, different subsets of variables may be dominant in each of the different clusters.

Alternatively, even when clustering genes on a relatively small number of samples, we may wish to cluster on only a subset of the samples if those samples correspond, say, to a particular group of experimental conditions. Thus we would want many "layers" of clustering based on different (and possibly overlapping) subsets of the tissue samples, with genes which are clustered together in one layer not necessarily together in another. Additive two-way analysis of variance (ANOVA) models for this purpose, termed plaid models for the rectangular blocking they suggest on the gene expression data matrix, were introduced by Lazzeroni and Owen [34].

Broadly speaking, there are two classes of clustering methods: hierarchical methods and optimisation methods. The former sequentially aggregates objects into clusters or sequentially split a large cluster into smaller ones, while the latter seeks those clusters which optimise some overall measure of clustering quality. We briefly summarise the methods below. Different algorithms are based on different measures of dissimilarity, and on different criteria determining how good a proposed cluster structure is. These differences naturally lead to different cluster structures. Put another way, such differences lead to different definitions of *what* a cluster is. A consequence of this is that one should decide what one means by a cluster before one chooses a method. The k-means algorithm described below will be good at finding compact spherical clusters of similar sizes, while the single-link algorithm is able to identify elongated sausage-shaped clusters. Which is appropriate depends on what sort of structure one is seeking. Merely because cluster analysis is an exploratory tool does not mean that one can apply it without thinking.

Hierarchical methods

Hierarchical clustering methods give rise to a sequence of nested partitions, meaning the intersection of a set in the partition at one level of the hierarchy with a set of the partition at a higher level of the hierarchy will always be equal to either the set from the lower level or the empty set. The hierarchy can thus be graphically represented by a tree. Typically this sequence will be as long as the number of observations (genes), so that level k of the hierarchy has exactly k clusters and the partition at level $k-1$ can be recovered by merging two of the sets in level k. In this case, the hierarchy can be represented by a binary tree, known as a "dendrogram." Usually the vertical scale of a dendrogram represents the distance between the two merged clusters at each level.

Methods to obtain cluster hierarchies are either top-down approaches, known as divisive algorithms, where one begins with a large cluster containing all the observations and successively divides it into finer partitions, or more commonly bottom-up, agglomerative algorithms, where one begins with each observation in its own cluster and successively merge the closest clusters until one large cluster remains. Agglomerative algorithms dominate the clustering literature because of the greatly reduced search space compared to divisive algorithms, the former usually requiring only $O(n^2)$ or at worst $O(n^3)$ calculations, whilst without reformulation performing the first stage of the latter alone requires $2^{n-1}-1$ calculations. This is reflected by the appearance of early versions of agglomerative hierarchical algorithms in the ecological and taxonomic literature as much as 50 years ago. To make divisive schemes feasible, monothetic approaches can be adopted, in which the possible splits are restricted to thresholds on single variables—in the same manner as the standard CART tree algorithm (Breiman et al [35]). Alternatively, at each stage the cluster with largest diameter can be "splintered" through allocating its largest outlier to a new cluster and relocating the remaining cluster members to whichever of the old and new clusters is closest, as in the *Diana* algorithm of Kaufman and Rousseeuw [36] which has been

implemented in the statistical programming language R. It has been suggested that an advantage of divisive methods is that they begin with the large structure in the data, again as in CART with its root split, but we have seen no examples to convince us that agglomerative methods are not equally enlightening.

Having selected an appropriate distance metric between observations, this needs to be translated into a "linkage metric" between clusters. In model-based clustering this again follows immediately, but otherwise natural choices are single link, complete link, or average link.

Single-link (or nearest neighbour) clustering defines the distance between two clusters as the distance between the two closest objects, one from each cluster (Sokal and Sneath [37] Jardine and Sibson [38]). A unique merit of the single-link method is that when one makes a choice between two equal intercluster distances for a merger, it will be followed by a merger corresponding to the other distance, which gives the method a certain type of robustness to small perturbations of the distances. Single-link clustering is susceptible to chaining: the tendency for a few points lying between two clusters to cause them to be joined. Whether this really is a weakness depends on what the aim is—on what one means by a "cluster." In general, if different objects are thought to be examples of the same kind of thing, but drawn at different stages of some developmental process, then perhaps one would want them to be assigned to the same cluster.

Complete-link (or furthest neighbour) clustering defines the distance between two clusters as the distance between the two furthest objects, one from each cluster. It is obvious that this will tend to lead to groups which have similar diameters, so that the method is especially valuable for dissection applications. Of course, if there are natural groups with very different diameters in the data, the smallest of these may well be merged before the large ones have been put together. We repeat, it all depends on one's aims and on what one means by a cluster.

Average-link (or centroid) clustering defines the distance between two clusters as the distance between the centroids of the two clusters. If the two clusters are of very different sizes, then the cluster that would result from their merger would maintain much of the characteristics of the larger cluster; if this is deemed undesirable, median cluster analysis which gives equal weighting to each cluster can be used.

Lance and Williams [39] present a simple linear system as a unifying framework for these different linkage measures.

After performing hierarchical clustering there remains the issue of choosing the number of clusters. In model-based clustering, this selection can be made using a model choice criterion such as Bayesian information criterion (Schwarz [40]) or in a Bayesian setting with prior distributions on model parameters, choosing the clustering which maximises marginal posterior probability. Otherwise, less formal procedures such as examining the dendrogram for a natural cut off or satisfying a predetermined upper bound on all within-group sums of squares are adopted.

Optimal partitioning methods

Perhaps more in tune with statistical ideas are direct partitioning techniques. These produce just a single "optimum" clustering of the observations rather than a hierarchy, meaning one must first state how many clusters there should be. In dissection analysis the number of groups is chosen by the investigator, but in cluster analysis the aim is to discover the naturally occurring groups, so some method is needed to compare solutions with different numbers of groups as discussed above at the end of hierarchical clustering.

For a fixed number of clusters k, a partitioning method seeks to optimise a clustering criterion; note, however, that the fact that no hierarchy is involved means that one may not be able to split a cluster in the solution with k clusters to produce the $k + 1$ cluster solution, and thus care must be taken in choosing a good starting point. Although, in principle, all one has to do is search over all possible allocations of objects to classes, seeking that particular allocation which optimises the clustering criterion, in practice there are normally far too many such possible allocations, so some heuristic search strategy must be adopted. Often, having selected an initial clustering, a search algorithm is used to iteratively relocate observations to different clusters until no gain can be made in the clustering criterion value.

The most commonly used partitioning method is "k-means" clustering (Lloyd [41] and MacQueen [42]). k-means clustering seeks to minimise the average squared distance between observations and their cluster centroid. This strategy can be initiated by specifying k centroids perhaps independently from the data, assigning each datum to the closest centroid, then recomputing the cluster centroids, reassigning each datum, and so on. Closely related to k-means clustering is the method of self-organising maps (SOM) (Kohonen [43]); these differ in also having prespecified geometric locations on which the clusters lie, such as points on a grid, and the clusters are iteratively updated in such a way that clusters close to each other in location tend to be relatively similar to one another. More generally, these optimisation methods usually involve minimising or maximising a criterion based on functions of the within-group ($\mathbf{W}$) and between-group ($\mathbf{B}$) (or, equivalently, the total $\mathbf{T}$) sum of squares and cross products matrix familiar from multivariate ANOVA. In fact, k-means clustering minimises trace ($\mathbf{W}$). Other common alternatives are minimising det ($\mathbf{W}$) and maximising trace ($\mathbf{BW}^{-1}$). For more details see Everitt [44].

Model-based partitioning methods are essentially mixture decomposition methods. Most commonly, mixtures of normal distributions are assumed, so that each cluster is characterised by an unknown mean and covariance matrix pair. Notable works in this area include Wolfe [21], Richardson and Green [45], and Fraley and Raftery

[20, 46, 47]. The authors of the latter provide the accompanying free software MCLUST, which uses the EM algorithm for parameter estimation to avoid the Markov chain Monte Carlo (MCMC) techniques required in Bayesian method of Richardson and Green [48] and Bensmail et al [19]. It should be remarked that when it comes to parameter estimation in mixture modelling, one has to be careful of the nonidentifiability of the mixture component labels; to get around this problem order constraints are placed on the parameters, often artificial, so that only a unique permutation of the component labels is supported.

Gene expression clustering

There are many instances of reportedly successful applications of both hierarchical clustering and partitioning techniques in gene expression analyses. This section illustrates the diversity of techniques which have been used.

Eisen et al [24] used agglomerative hierarchical clustering with their uncentred correlation-based dissimilarity metric as described above for growth time-course microarray data from budding yeast. This approach has since been followed in similar studies by Chu et al [49], Spellman et al [50], Iyer et al [51], Perou et al [52] and Nielsen et al [4]. Alternatively, Wen et al [53] used Euclidean hierarchical clustering on vectors with the time series of expression levels for each concatenated with the slopes between them to take into account offset but parallel patterns.

Turning to nonmodel-based partitioning methods, SOMs have been favoured; Tamayo et al [54] used SOMs for clustering of different time series of gene expression data. Similar approaches have also been used by Golub et al [1] for cancer tissue class discovery and prediction and Kasturi et al [55] for gene expression time series, where the latter first normalises the data to allow the use of Kullback-Leibler divergence as the distance metric. Tavazoie et al [56] represented expression time series in T-dimensional space and used the k-means clustering algorithm.

To find more subtle cluster structures, many model-based variations have been developed beyond these generic methods. This has been especially beneficial in the context of time series of gene expression samples. Ramoni et al [57] modelled gene expression time series with autoregressive processes, providing the accompanying free software CAGED. Luan and Li [58] clustered gene expression time series with mixed effects with B-splines; Bar-Joseph et al [59] used cubic splines for each gene with spline coefficients constrained to be similar for genes in the same cluster. They also used a time warping algorithm to align time series with similar expression profiles in different phases. Wakefield et al [60] performed clustering using a full MCMC Bayesian approach, with a basis function representation for the expression time series incorporating random effects. Yeung et al [61] used the mixture of normal distributions software MCLUST of Fraley and Raftery [20, 46, 47] for a variety of real and synthetic gene expression data sets, some time indexed. Pan et al [62] used the same model as MCLUST but on a two-sample t-statistic of differential expression for each gene rather than the full gene expression data matrix. Medvedovic and Sivaganesan [63] used the Gibbs sampling methods of Neal [64] for Dirichlet process mixture models to give a Bayesian version. Alon et al [65] used a divisive algorithm iteratively fitting two Gaussians at each stage with self-consistent equations. Heard et al [66] used a mixture of Gaussian processes with basis function representations for clustering of gene expression time series, with a conjugate model removing the need for MCMC.

Graphical models have also been attempted. Ben-Dor et al [67] gave two alternative graphical model-based clustering algorithms, clustering genes on a similarity matrix, PCC and CAST. Zhou et al [8] connected genes with highly correlated gene expression in a graphical model and clustered genes through a shortest-path analysis identifying "transitive genes." Dobra et al [68] attempted to actually model the whole covariance structure of the genes using Gaussian graphical models.

Instead of working on the raw gene expression matrix (genes $\times$ arrays), Alter et al [69] used singular value decomposition (cf principal component analysis) to analyse microarray data in the reduced diagonalised "eigengenes"$\times$"eigenarrays" space and filter out the eigengenes or eigenarrays inferred to represent experimental noise. Clustering in this new space was performed by Holter et al [70, 71] using standard hierarchical techniques; by Hastie et al [11] using "gene shaving," which identifies subsets of the genes with coherent expression patterns and large variations across samples; and by Wall et al [72] using thresholding on the magnitude of the elements of the left singular vectors (gene coefficient vectors), this thresholding enabling genes inhibited or promoted by the same transcription regulator to be clustered together. Clustering on principal components using k-means, Euclidean hierarchical average link and CAST was tested by Yeung and Ruzzo [73] and showed no benefits over clustering on the raw data.

Heyer et al [74] devised QT-clustering. There, for robustness to outliers, the *jackknife correlation* between two gene expression vectors is taken as the minimum of the correlation between the whole of both vectors or the correlation of the two vectors with any single component deleted. Clusters are then iteratively generated, with each made to be as large as possible subject to a threshold on the diameter of the cluster.

It will be apparent that much of the above hinges on how the distance between profiles is measured. Indeed, in general, different ways of measuring distance will lead to different solutions. This leads on to the question of how to assess the performance of different methods. In general, since most of these problems are fundamentally exploratory, there is no ideal answer to this. Datta and Datta [75] compared k-means, hierarchical (raw and partial least squares regression), MCLUST,

Diana and Fanny (a fuzzy k-means algorithm, see Kaufman and Rousseeuw [36]) for real temporal and replicate microarray gene expression data, scoring each method using measures of cluster overlap and distance, and overall favoured Diana. Yeung et al [76] compared k-means clustering, CAST, single-, average- and complete-link hierarchical clustering, and totally random clustering for both simulated and real gene expression data, scoring each method predictively using a jackknife procedure to obtain an adjusted "figure of merit," and favoured k-means and CAST. Gibbons and Roth [77] compared k-means, SOMs, and hierarchical clustering of real temporal and replicate microarray gene expression data, using a figure of merit which scores against random assignment, and favoured k-means and SOMs. For single method cluster validation, Li and Wong [78] used bootstrap sampling to check cluster membership robustness after hierarchical clustering.

In general, different clustering methods may yield different clusters. This is hardly surprising, given that they define what is meant by a cluster in different ways. It is true that if there is a very strong clustering in the data, one might expect consistency among the results, but it is less true that differences in the discovered cluster structure means that there is no cluster structure.

PATTERN DISCOVERY

Cluster analysis partitions a data set and, by implication, the space in which the data are embedded. All data points, and all possible data points, are assigned to an element of the partition. Often, however, one does not wish to make such grand sweeping statements. Often one merely seeks to find *localised* subsets of objects, in the sense that a set of objects are behaving in an unexpectedly similar way, regardless of the remainder of the objects. In the context of gene expression data, this would mean that amongst the mass of genes, each with their own expression profile, a (possibly) small number had unusually similar profiles. (As mentioned earlier, this idea can be generalised—one might be interested in detecting negatively correlated expression profiles—but we will not discuss such generalisations here.) In the context of nucleotide sequencing, it would mean that interest lay in identifying sequences which were very similar, without any preconceptions about what sort of sequence one was searching for. In both of these examples, one begins, as one does in cluster analysis, with the concept of a distance between elements (expression profiles or nucleotide sequences), but here, instead of using this distance to partition the data space, one merely uses it to find locally dense regions of the data space. Note that, in these two examples, the distance measures used are very different: classic multivariate distance measures (Euclidean distance being the most familiar) can be used in the first case, but the second case requires measures of distances between sequences or strings of symbols, such as the Levenshtein distance (Levenshtein [79]). In such situations, a natural way to define distance is in terms of the number of edit operations needed to make one string identical to the other. If edit operations are defined as being one of insertion, deletion, or substitution, then the Levenshtein (or "edit" or "sequence") distance between two strings is the minimum number of such operations needed to convert from one string to the other. A distance of 0 corresponds to an exact match. Since an optimisation is involved here, to find the minimum, many distance measures for strings use ideas of dynamic programming.

One stream of work aiming at detecting locally dense accumulations of data points goes under the name of "bump hunting" (eg, Silverman [80] and Harezlak [81]). Early work concentrated on unidimensional problems, but this has now been extended to multiple dimensions. An example is the PRIM algorithm of Friedman and Fisher [82], which embeds the ideas in a more general framework. Work which is intrinsically multivariate includes the PEAKER algorithm described in [83, 84], which identifies those data point locations which have a higher estimated data probability density than all local neighbours.

Although we have described the exercise as being one of finding localised groups of objects in the data set, in fact the aim is really typically one of inference. For example, the question is not really whether some particular expression profiles in the database are surprisingly similar, but whether these represent real underlying similarities between genes. There is thus an inferential aspect involved, which allows for measurement error and other random aspects of the process producing the data to make statements about the underlying structure. The key question implicit in this inferential aspect is whether the configuration could have arisen by chance, or whether it is real in the sense that it reflects an unusually high local density in the distribution of possible profiles. Sometimes the unusually high local probability densities are called "patterns" (eg, Hand et al [10]).

In order to make a statement about whether a configuration of a few data points is *unexpectedly* dense, one needs to have some probability model with which to compare it. In spatial epidemiology this model is based on the overall population distribution, so that, for example, one can test whether the proportion of cases of illness is unexpectedly high in a particular local region. In general, however, in bioinformatics applications such background information may not be available. DuMouchel [85] gives a particularly telling example of this in the context of adverse drug reactions, where there is no information about the overall number of times each drug has been prescribed. In such cases, one has to make reasonable assumptions about the background model. This is not necessarily problematic: the aim is, after all, an exploratory one. A basic form of background model is an inhomogeneous Poisson process, with the local intensity being based on any information one does have. Since, typically, more than one variable is involved, background models based on independence or information about known

relationships (such as time order) between the variables are often used.

From an inferential perspective, the key issue is one of multiplicity. With a large data space, perhaps with many observations, there is considerable opportunity for a large number of local maxima of an estimated density function. Deciding which of these maxima are genuine and which are attributable to chance is a nontrivial problem, and one which has been of concern in more general data mining contexts. Traditional statistical approaches to the multiplicity problem focus on controlling familywise error rate, setting a limit on the proportion of cases where there is no underlying distributional structure which are detected as significant. The consequence is that only the largest probability peaks exceed the chosen threshold. Benjamini and Hochberg [86], however (although in a rather different context), suggested controlling the (expected value of the) proportion of structures detected as significant where no real structure existed. This does not require so great a sacrifice of power for the individual tests. More generally, there is an accumulating body of statistical work in the area of *scan statistics* (eg, Glaz et al [87]). The intuitive idea here is that one scans a window over the data space, seeking positions where some function of the data within the window (eg, in our case, a count of the data points) exceeds some critical value. To date most of this work has focused on low-dimensional cases.

Compared to cluster analysis, pattern discovery is a relatively new area of investigation, but, like cluster analysis, the ideas have been developed by several different intellectual communities for different problem domains contemporaneously. These include speech recognition, text processing (which, of course, has received a dramatic boost with the web), real-time correction of keyboard entry errors, technical analysis ("chartism") in tracking stock prices, association analysis (in data mining, including its subdiscipline of market basket analysis), configural frequency analysis, and other areas. Ideas developed in one area can often be ported across to others, and, in particular, to bioinformatics—just as cluster analysis has been. There are also areas of statistics, which, although they have their own special emphases, are closely related, such as outlier detection.

Bump hunting, pattern discovery, or peak detection methods seem to have been applied relatively rarely in the analysis of microarray data to date, and yet in many problems such tools are arguably more appropriate than cluster analysis. In particular, in cases where the aim is to group the genes on the basis of their time or experimental condition expression profiles, many, perhaps most, of the genes will be doing nothing of interest. Including those in the partitioning is at best pointless and at worst may be misleading.

CONCLUSION

Cluster analysis with gene expression data has its own aspects, perhaps notably that of high dimensionality and low number of cases for some problems. However, in other ways this domain avoids issues which are important for other applications. The question of scaling the variables has been mentioned above: typically gene expression variables are commensurate. Choice of variables is a critical problem when one is trying to classify cells or samples, on the basis of the genes (and hence with a very large number of variables) but unimportant when one is trying to classify genes themselves, perhaps on the basis of very few expression conditions. This is a crucial issue, since cluster structure is always in the context of the variables chosen to describe the objects.

The most important point to bear in mind when considering using cluster analysis is that different methods have different (often implicit) definitions of what is meant by a cluster. If one is searching for compact spherical structures in the database, for example, one should not use a method which is likely to throw up long attenuated structures—and vice versa. Of course, in a completely exploratory situation, one can argue that any kind of structure could be of interest. This is true and provides a case for using multiple different methods (with multiple different distance measures) in the hope that some interesting structure may be found. In general, however, one does better by constraining the exploration in the light of what one already knows or believes likely to be the case: as Louis Pasteur said, "chance favours only the prepared mind."

Even more generally than the question of whether researchers have always used the appropriate method of cluster analysis when analysing their microarray data is the question of whether *any* form of cluster analysis is appropriate. Cluster analysis is a partitioning tool, assigning each of the data points to a unique (in general) cluster. Often, however, much, perhaps most, of the data points are uninteresting, with concern only being with particular local regions of the data space. In this context, pattern discovery methods in particular seem relevant to the analysis of microarray data. These tools identify subgroups of objects (eg, genes) which have similar profiles, regardless of the profile shapes of the other objects.

ACKNOWLEDGMENT

The work of Nicholas Heard described in this paper was funded by the Wellcome Trust grant number 065822.

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531–537.

[2] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000;403(6769):503–511.

[3] Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature.* 2000;406(6795):536–540.
[4] Nielsen TO, West RB, Linn SC, et al. Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet.* 2002;359(9314):1301–1307.
[5] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA.* 2002;99(10):6567–6572.
[6] Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *J Roy Statist Soc Ser B.* 2002;64(4):717–736.
[7] Dhillon IS, Marcotte EM, Roshan U. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*. 2003;19:1612–1619.
[8] Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA.* 2002;99(20):12783–12788.
[9] Kendall MG. *Multivariate Analysis.* 2nd ed. New York, NY: Macmillan; 1980.
[10] Hand DJ, Adams NM, Bolton RJ, eds. *Pattern Detection and Discovery.* New York, NY: Springer; 2002.
[11] Hastie T, Tibshirani R, Eisen MB, et al. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):Research0003.
[12] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist Sinica.* 2002;12(1):111–139.
[13] Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002;30(4):e15.
[14] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4(2):249–264.
[15] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003;19(2):185–193.
[16] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 2001;17(6):520–525.
[17] Turkheimer FE, Duke DC, Moran LB, Graeber MB. Wavelet analysis of gene expression (WAGE). In: *Proceedings of the IEEE International Symposium on Biomedical Imaging: Macro to Nano. Vol* 2. Arlington, Va; 2004:1183–1186.
[18] Bock H. Probabilistic models in cluster analysis. *Comput Stat Data An.* 1996;23:5–28.
[19] Bensmail H, Celeux G, Raftery AE, Robert CP. Inference in model-based cluster analysis. *Stat Comput.* 1997;7:1–10.
[20] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 2002;97(458):611–631.
[21] Wolfe JH. *A Computer Program for Maximum-Likelihood Analysis of Types.* San Diego, Calif: US Naval Personnel Research Activity; 1965. USNPRA Technical Bulletin 65-15.
[22] Gower JC. Measures of similarity, dissimilarity, and distance. In: Kotz S, Johnson NL, Read CB, eds. *Encyclopedia of Statistical Sciences. Vol* 5. New York, NY: John Wiley & Sons; 1985:397–405.
[23] Gordon AD. *Classification*. 2nd ed. Boca Raton, Fla: Chapman and Hall/CRC; 1999.
[24] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression *Proc Natl Acad Sci USA.* 1998;95(25):14863–14868.
[25] Milligan GW. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika.* 1980;45:325–342.
[26] DeSarbo WS, Mahajan V. Constrained classification: the use of a priori information in cluster analysis. *Psychometrika*. 1984;49:187–215.
[27] Fowlkes EB, Gnanadesikan R, Kettenring JR. Variable selection in clustering. *J Classification.* 1988;5(2):205–228.
[28] De Soete G, DeSarbo WS, Carroll JD. Optimal variable weighting for hierarchical clustering: an alternating least squares algorithm. *J Classification.* 1985;2:173–192.
[29] De Soete G. Optimal variable weighting for ultrametric and additive tree fitting. *Qual Quant.* 1986;20:169–180.
[30] De Soete G. OVWTRE: a program for optimal variable weighting for ultrametric and additive tree fitting. *J Classification.* 1988;5:101–104.
[31] Van Buuren S, Heiser W. Clustering n objects into k groups under optimal scaling of variables. *Psychometrika.* 1989;54:699–706.
[32] Brusco MJ, Cradit JD. A variable selection heuristic for k-means clustering. *Psychometrika*. 2001;66:249–270.
[33] Friedman JH, Meulman JJ. Clustering objects on subsets of attributes (with discussion). *J Roy Statist Soc Ser B.* 2004;66(4):815–849.
[34] Lazzeroni LC, Owen A. Plaid models for gene expression data. *Statist Sinica.* 2002;12(1):61–86.
[35] Breiman L, Friedman JH, Olshen R, Stone CJ. *Classification and Regression Trees.* Belmont, Calif: Wadsworth Advanced Books and Software; 1984.
[36] Kaufman L, Rousseeuw PJ. *Finding Groups in Data. An Introduction to Cluster Analysis.* New York, NY: John Wiley & Sons; 1990.
[37] Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. San Francisco, Calif: WH Freeman; 1963.
[38] Jardine N, Sibson R. *Mathematical Taxonomy*. London, UK: Wiley; 1971.

[39] Lance GN, Williams WT. A general theory of classification sorting strategies. I. Hierarchical systems. *Comput J*. 1967;9:373–380.
[40] Schwarz G. Estimating the dimension of a model. *Ann Statist.* 1978;6(2):461–464.
[41] Lloyd SP. *Least Squares Quantization in PCM.* Murray Hill, NJ: Bell Laboratories; 1957. Internal Technical Report. Published in IEEE Transactions on Information Theory.
[42] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, eds. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol 1 of Statistics*. Berkeley, Calif: University of California Press; 1967:281–297.
[43] Kohonen T. *Self-organizing Maps.* Berlin, Germany: Springer; 1995.
[44] Everitt BS. *Cluster Analysis.* 3rd ed. London, UK: Edward Arnold; 1993.
[45] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components. *J Roy Statist Soc Ser B.* 1997;59:731–792.
[46] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J.* 1998;41(8):578–588.
[47] Fraley C, Raftery AE. MCLUST: software for model-based cluster analysis. *J Classification.* 1999;16(2):297–306.
[48] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J Roy Statist Soc Ser B.* 1997;59(4):731–792.
[49] Chu S, DeRisi J, Eisen MB, et al. The transcriptional program of sporulation in budding yeast. *Science.* 1998;282(5389):699–705.
[50] Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell.* 1998;9(12):3273–3297.
[51] Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum. *Science.* 1999;283(5398):83–87.
[52] Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA.* 1999;96(16):9212–9217.
[53] Wen X, Fuhrman S, Michaels GS, et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA.* 1998;95(1):334–339.
[54] Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA.* 1999;96(6):2907–2912.
[55] Kasturi J, Acharya R, Ramanathan M. An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics.* 2003;19(4):449–458.
[56] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet.* 1999;22(3):281–285.
[57] Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci USA.* 2002;99(14):9121–9126.
[58] Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics.* 2003;19(4):474–482.
[59] Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I. A new approach to analyzing gene expression time series data. In: *Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB '02).* Washington, DC; 2002:39–48.
[60] Wakefield J, Zhou C, Self S. Modelling gene expression over time: curve clustering with informative prior distributions. In: Bernardo JM, Bayarri MJ, Berger JO, Heckerman D, Smith AFM, West M, eds. *Proceedings of the 7th Valencia International Meeting. Vol 7 of Bayesian Statistics*. New York, NY: The Clarendon Press, Oxford University Press; 2003:721–732.
[61] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics.* 2001;17(10):977–987.
[62] Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. *Genome Biol.* 2002;3(2):Research0009.
[63] Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics.* 2002;18(9):1194–1206.
[64] Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat.* 2000;9(2):249–265.
[65] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA.* 1999;96(12):6745–6750.
[66] Heard NA, Holmes CC, Stephens DA. *A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves*. London, UK: Imperial College; 2004. Technical Report.
[67] Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol.* 1999;6(3–4):281–297.
[68] Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. *J Multivariate Anal*. 2004;90(1):196–212.
[69] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA.* 2000;97(18):10101–10106.

[70] Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci USA.* 2000;97(15):8409–8414.
[71] Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. *Proc Natl Acad Sci USA.* 2001;98(4):1693–1698.
[72] Wall ME, Dyck PA, Brettin TS. SVDMAN—singular value decomposition analysis of microarray data. *Bioinformatics.* 2001;17(6):566–568.
[73] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics.* 2001;17(9):763–774.
[74] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 1999;9(11):1106–1115.
[75] Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics.* 2003;19(4):459–466.
[76] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics.* 2001;17(4):309–318.
[77] Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* 2002;12(10):1574–1581.
[78] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2001;2(8):Research0032.
[79] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory.* 1966;10(8):707–710.
[80] Silverman BW. Using kernel density estimates to investigate multimodality. *J Roy Statist Soc Ser B.* 1981;43(1):97–99.
[81] Harezlak J. *Bump Hunting Revisited* [master's thesis]. Vancouver, BC, Canada: Department of Statistics, University of British Columbia; 1998.
[82] Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). *Stat Comput.* 1999;9:123–162.
[83] Adams NM, Hand DJ, Till RJ. Mining for classes and patterns in behavioural data. *J Opl Res Soc.* 2001;52:1017–1024.
[84] Bolton RJ, Hand DJ, Crowder MJ. Significance tests for unsupervised pattern discovery in large continuous multivariate data sets. *Comput Statist Data Anal.* 2004;46(1):57–79.
[85] DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system (with discussion). *Am Stat.* 1999;53:177–202.
[86] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B.* 1995;57(1):289–300.
[87] Glaz J, Naus J, Wallenstein S. *Scan Statistics.* New York, NY: Springer; 2001.