Journal of Biomedicine and Biotechnology • 2003:4 (2003) 217–230 • PII. S1110724303209207 • http://jbb.hindawi.com

REVIEW ARTICLE

Postgenomics: Proteomics and Bioinformatics in Cancer Research

Halima Bensmail[1] and Abdelali Haoudi[2]*

[1]*Department of Statistics, University of Tennessee, Knoxville, TN 37996, USA*
[2]*Department of Microbiology and Molecular Cell Biology and the Virginia Prostate Center,*
Eastern Virginia Medical School, 700 West Olney Road, Norfolk, VA 23501, USA

Received 26 September 2002; revised 30 November 2002; accepted 30 November 2002

Now that the human genome is completed, the characterization of the proteins encoded by the sequence remains a challenging task. The study of the complete protein complement of the genome, the "proteome," referred to as proteomics, will be essential if new therapeutic drugs and new disease biomarkers for early diagnosis are to be developed. Research efforts are already underway to develop the technology necessary to compare the specific protein profiles of diseased versus nondiseased states. These technologies provide a 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 identification of metabolic disorders. Here, we present an overview of the current status and future research approaches in defining the cancer cell's proteome in combination with different bioinformatics and computational biology tools toward a better understanding of health and disease.

TECHNOLOGIES FOR PROTEOMICS

2D gel electrophoresis

Two-dimensional gel electrophoresis (2DE) is by far the most widely used tool in proteomics approaches for more than 25 years [1]. This technique involves the separation of complex mixtures of proteins first on the basis of isoelectric point (pI) using isoelectric focusing (IEF) and then in a second dimension based on molecular mass. The proteins are separated by migration in a polyacrylamide gel. By use of different gel staining techniques such as silver staining [2], Coomassie blue stain, fluorescent dyes [3], or radiolabels, few thousands proteins can be visualized on a single gel. Fluorescent dyes are being developed to overcome some of the drawbacks of silver staining in making the protein samples more amenable to mass spectrometry [4, 5]. Stained gels can then be scanned at different resolutions with laser densitometers, fluorescent imager, or other device. The data can be analyzed with software such as PDQuest by Bio-Rad Laboratories (Hercules, Calif, USA) [6], Melanie 3 by GeneBio (Geneva, Switzerland), Imagemaster 2D Elite by Amersham Biosciences, and DeCyder 2D Analysis by Amersham Biosciences (Buckinghamshire, UK) [7]. Ratio analysis is used to detect quantitative changes in proteins between two samples. 2DE is currently being adapted to high-throughput platforms [8]. For setting up a high-throughput environment for proteome analysis, it is essential that the 2D gel image analysis software supports robust database tools for sorting images, as well as data from spot analysis, quantification, and identification.

ProteinChips

While proteomics has become almost synonymous with 2D gel electrophoresis, there is a variety of new methods for proteome analysis. Unique ionization techniques, such as electrospray ionization and matrix-assisted laser desorption-ionization (MALDI), have facilitated the characterization of proteins by mass spectrometry (MS) [9, 10]. These techniques have enabled the transfer of the proteins into the gas phase, making it conducive for their analysis in the mass spectrometer. Typically, sequence-specific proteases are used to break up the proteins into peptides that are coprecipitated with a light-absorbing matrix such as dihydroxy benzoic acid. The peptides are then subjected to short pulses of ultraviolet radiation under reduced pressure. Some of the peptides are ionized and accelerated in an electric field and subsequently turned back through an energy correction device [11]. Peptide mass is derived through a time-of-flight (TOF) measurement of the elapsed time from acceleration-to-field free drift or through a quadrupole detector. A peptide mass map is generated with the sensitivity to detect molecules at a few parts per million. Hence a spectrum is generated with the molecular mass of individual peptides, which are used to search databases to find matching proteins. A minimum of three peptide molecular weights is necessary to minimize false-positive matches.

The principle behind peptide mass mapping is the matching of experimentally generated peptides with those determined for each entry in a sequence. The alternative process of ionization, through the electrospray ionization, involves dispersion of the sample through a capillary device at high voltage [11]. The charged peptides pass through a mass spectrometer under reduced pressure and are separated according to their mass-to-charge ratios through electric fields. After separation through 2DE, digested peptide samples can be delivered to the mass spectrometer through a "nanoelectrospray" or directly from a liquid chromatography column (liquid chromatography-MS), allowing for real-time sequencing and identification of proteins. Recent developments have led to the MALDI quadrupole TOF instrument, which combines peptide mapping with peptide sequencing approach [12, 13, 14]. An important feature of tandem MS (MS-MS) analysis is the ability to accurately identify posttranslational modifications, such as phosphorylation and glycosylation, through the measurement of mass shifts.

Another MS-based proteinChip technology, surface-enhanced laser desorption-ionization time of flight mass spectrometry (SELDI-TOF-MS), has been successfully used to detect several disease-associated proteins in complex biological specimens, such as cell lysates, seminal plasma, and serum [15, 16, 17]. Surface-enhanced laser desorption-ionization (SELDI) is an affinity-based MS method in which proteins are selectively adsorbed to a chemically modified surface, and impurities are removed by washing with buffer. The use of several different chromatographic arrays and wash conditions enables high-speed, high-resolution chromatographic separations [14].

Other technologies

Arrays of peptides and proteins provide another biochip strategy for parallel protein analysis. Protein assays using ordered arrays have been explored through the development of multipin synthesis [18]. Arrays of clones from phage-display libraries can be probed with antigen-coated filters for high-throughput antibody screening [19]. Proteins covalently attached to glass slides through aldehyde-containing silane reagents have been used to detect protein-protein interactions, enzymatic targets, and protein small molecule interactions [20]. Other methods of generating protein microarrays are by printing the proteins (ie, purified proteins, recombinant proteins, and crude mixtures) or antibodies using a robotic arrayer and a coated microscope slide in an ordered array. Protein solutions to be measured are labeled by covalent linkage of a fluorescent dye to the amino groups on the proteins [21]. Protein arrays consisting of immobilized proteins from pure populations of microdissected cells have been used to identify and track cancer progression. Although protein arrays hold considerable promise for functional proteomics and expression profiling for monitoring a disease state, certain limitations need to be overcome. These include the development of high-throughput technologies to express and purify proteins and the generation of large sets of well-characterized antibodies. Generating protein and antibody arrays is more costly and labor-intensive relative to DNA arrays. Nevertheless, the availability of large antibody arrays would enhance the discovery of differential biomarkers in nondiseased and cancer tissue [22].

Tissue arrays have been developed for high-throughput molecular profiling of tumor specimens [23]. Arrays are generated by robotic punching out of small cylinders (0.6 mm × 3–4 mm high) of tissue from thousands of individual tumor specimens embedded in paraffin to array them in a paraffin block. Tissue from as many as 600 specimens can be represented in a single "master" paraffin block. By use of serial sections of the tissue array, tumors can be analyzed in parallel by immunohistochemistry, fluorescence in situ hybridization, and RNA-RNA in situ hybridization. Tissue arrays have applications in the simultaneous analysis of tumors from many different patients at different stages of disease. Disadvantages of this technique are that a single core is not representative because of tumor heterogeneity and uncertainty of antigen stability on long-term storage of the array. Hoos et al [24] demonstrated that using triplicate cores per tumor led to lower numbers of lost cases and lower nonconcordance with typical full sections relative to one or two cores per tumor. Camp et al [25] found no antigenic loss after storage of an array for 3 months. Validation of tissue microarrays is currently ongoing in breast and prostate cancers and will undoubtedly help in protein expression profiling [23, 25, 26]. A major advantage of this technology is that expression profiles can be correlated with outcomes from large cohorts in a matter of few days.

PROTEOMICS IN CANCER RESEARCH

Cancer proteomics encompasses the identification and quantitative analysis of differentially expressed proteins relative to healthy tissue counterparts at different stages of disease, from preneoplasia to neoplasia. Proteomic technologies can also be used to identify markers for cancer diagnosis, to monitor disease progression, and to identify therapeutic targets. Proteomics is valuable in the discovery of biomarkers because the proteome reflects both the intrinsic genetic program of the cell and the impact of its immediate environment. Protein expression and function are subject to modulation through transcription as well as through posttranscriptional and posttranslational events. More than one RNA can result from one gene through a process of differential splicing. Additionally, there are more than 200 posttranslation modifications that proteins could undergo, that affect function, protein-protein and nuclide-protein interaction, stability, targeting, half-life, and so on [27], all contributing to a potentially large number of protein products from one gene. At the protein level, distinct changes occur during the transformation of a healthy cell into a neoplastic cell, ranging from altered expression, differential protein modification, and changes in specific activity, to aberrant localization, all of which may affect cellular function. Identifying and understanding these changes are the underlying themes in cancer proteomics. The deliverables include identification of biomarkers that have utility both for early detection and for determining of therapy.

Although proteomics traditionally dealt with quantitative analysis of protein expression, more recently, proteomics has been viewed to encompass the structural analysis of proteins [28]. Quantitative proteomics strives to investigate the changes in protein expression in different states, such as in healthy and diseased tissue or at different stages of the disease. This enables the identification of state- and stage-specific proteins. Structural proteomics attempts to uncover the structure of proteins and to unravel and map protein-protein interactions.

MS has been helpful in the analysis of proteins from cancer tissues. Screening for the multiple forms of the molecular chaperone 14-3-3 protein in healthy breast epithelial cells and breast carcinomas yielded a potential marker for the noncancerous cells [29]. The 14-3-3 form was observed to be strongly down regulated in primary breast carcinomas and breast cancer cell lines relative to healthy breast epithelial cells. This finding, in the light of the evidence that the gene for 14-3-3 was found silenced in breast cancer cells [30], implicates this protein as a tumor suppressor. Using a MALDI-MS system, Bergman et al [6] detected increases in the expressions of nuclear matrix, redox, and cytoskeletal proteins in breast carcinoma relative to benign tumors. Fibroadenoma exhibited an increase in the oncogene product DJ-1. Retinoic acid-binding protein, carbohydrate-binding protein, and certain lipoproteins were increased in ovarian carcinoma, whereas cathepsin D was increased in lung adenocarcinoma.

Imaging MS is a new technology for direct mapping and imaging of biomolecules present in tissue sections. For this system, frozen tissue sections or individual cells are mounted on a metal plate, coated with ultraviolet-absorbing matrix, and placed in the MS. With the use of an optical scanning raster over the tissue specimen and measurement of the peak intensities over thousands of spots, MS images are generated at specific mass values [31]. Stoeckli et al [32] used imaging MS to examine protein expression in sections of human glioblastoma and found increased expression of several proteins in the proliferating area compared with healthy tissue. Liquid chromatography—MS and tandem MS (MS-MS) were used to identify thymosin ß.4, a 4964-d protein found only in the outer proliferating zone of the tumor [32]. Imaging MS shows potential for several applications, including biomarker discovery, biomarker tissue localization, understanding of the molecular complexities of tumor cells, and intraoperative assessment of surgical margins of tumors.

SELDI, originally described by Hutchens and Yip [33], overcomes many of the problems associated with sample preparations inherent with MALDI-MS. The underlying principle in SELDI is surface-enhanced affinity capture through the use of specific probe surfaces or chips. This protein biochip is the counterpart of the array technology in the genomic field and also forms the platform for Ciphergen's ProteinChip array SELDI MS system [14]. A 2DE analysis separation is not necessary for SELDI analysis because it can bind protein molecules on the basis of its defined chip surfaces. Chips with broad binding properties, including immobilized metal affinity capture, and with biochemically characterized surfaces, such as antibodies and receptors, form the core of SELDI. This MS technology enables both biomarker discovery and protein profiling directly from the sample source without preprocessing. Sample volumes can be scaled down to as low as 0.5 μL, an advantage in cases in which sample volume is limiting. Once captured on the SELDI protein biochip array, proteins are detected through the ionization-desorption TOF-MS process. A retentate (proteins retained on the chip) map is generated in which the individual proteins are displayed as separate peaks on the basis of their mass and charge (m/z). Wright et al [15] demonstrated the utility of the ProteinChip SELDI-MS in identifying known markers of prostate cancer and in discovering potential markers either over- or underexpressed in prostate cancer cells and body fluids. SELDI analyses of cell lysates prepared from pure populations from microdissected surgical tissue specimens revealed differentially expressed proteins in the cancer cell lysate when compared with healthy cell lysates and with benign prostatic hyperplasia (BPH) and prostate intraepithelial neoplasia cell lysates [15]. SELDI is a method that provides protein profiles or patterns in a short period of time from a small starting sample, suggesting that molecular fingerprints may provide insights into changing protein expression from healthy to benign to premalignant to malignant lesions. This appears to be the case because distinct SELDI protein profiles for each cell and cancer type evaluated, including prostate, lung, ovarian, and breast cancer, have been described recently [34, 35]. After prefractionation, a SELDI profile of 30 dysregulated proteins was observed in seminal plasma from prostate cancer patients. One of the seminal plasma proteins detected by comparing the prostate cancer profiles with a BPH profile was identified as seminal basic protein, a proteolytic product of semenogelin I [14].

BIOINFORMATICS TOOLS

Bioinformatics tools are needed at all levels of proteomic analysis. The main databases serving as the targets for MS data searches are the expressed sequence tag and the protein sequence databases, which contain protein sequence information translated from DNA sequence data [11]. It is thought that virtually any protein that can be detected on a 2D gel can be identified through the expressed sequence tag database, which contains over 2 million cDNA sequences [36]. A modification of sequence-tag algorithms has been shown to locate peptides given the fact that the expressed sequence tags cover only a partial sequence of the protein [37].

Data mining for proteomics

A number of algorithms have been proposed for genomes-scale analysis of patterns of gene expression, including expressed sequence tags (ESTs) (simple expedient of counting), UniGene for gene indexes [38]. Going beyond expression data, efforts in proteomics can be expressed to fill in a more complete picture of posttranscriptional events and the overall protein content of cells. To address the large-in-scale data, this review addresses primarily those advances in recent years.

Concurrent to the development of the genome sequences for many organisms, MS has become a valuable technique for the rapid identification of proteins and is now a standard more sensitive and much faster alternative to the more traditional approaches to sequencing such as Edman degradation.

Due to the large array of data that is generated from a single analysis, it is essential to implement the use of algorithms that can detect expression patterns from such large volumes of data correlating to a given biological/pathological phenotype from multiple samples. It enables the identification of validated biomarkers correlating strongly to disease progression. This would not only classify the cancerous and noncancerous tissues according to their molecular profile but could also focus attention upon a relatively small number of molecules that might warrant further biochemical/molecular characterization to assess their suitability as potential therapeutic targets. Data screened is usually of large size and has about 100 000–120 000 variables.

Biologists are not prepared to handle the huge data produced by the proteins or DNA microarray projects or to use the "eye" to visualize and interpret the output, therefore to detect pattern, visualize, classify, and store the data, more sophisticated tools are needed. Bioinformatics has proved to be a powerful tool in the effective generation of primarily predictive proteomic data from analysis of DNA sequences. Proteomics studies applications and techniques, includes profiling expression patterns in response to various variables and conditions and time correlation analysis of protein expression.

Intelligent data mining facilities are essential if we are to prevent important results from being lost in the mass of information. The analysis of data can proceed with different levels. One level of differential analysis where genes are analyzed one by one independently of each other to detect changes in expression across different conditions. This is challenging due to the amount of noise involved and low repetition characteristic of microarray experiments. The next level of analysis involves visualizing and feature discovery. Basic statistical tools and statistical inferences include cluster analysis, Bayesian modeling, classification, and discrimination, neural networks, and graphical models. The basic idea behind those approaches is to visualize the correlations in the data to allow the data to be examined for similarity and detection of important expression patterns (principal component analysis) to learn (classification, neural networks, support vector machine), to predict (prediction, regression, regression tree), to detect feature discovery, and to test hypotheses regarding the number of distinct clusters contained within the data (hierarchical clustering, Bayesian clustering, *k*-means, mixture model with Gibbs sampler or EM algorithm).

These algorithms can quickly analyze gels to identify how a series of gels are related, for example, confirming separation of clusters into healthy (control), diseased, and treatments clusters, or perhaps pointing to the existence of a cluster which has not previously been considered, which is a population of cells exhibiting drug resistance [39, 40].

Principal component analysis

Principal component analysis (PCA) can be an effective method of identifying the most discriminating features in a data set. This technique usually involves finding two or three linear combinations of the original features that best summarize the types of variation in the data. If much of the variation is captured by these two or three most significant principal components, class membership of many data points can be observed. One may use the principal-component solution to the factor model for extracting factors (components). This is accomplished by the use of the principal-axis theorem, which says that for a gene-by-gene ($n \times n$) correlation matrix $\mathbf{R}$, there exists a rotation matrix $\mathbf{D}$ and diagonal matrix $\mathbf{\Lambda}$ such that $\mathbf{DRD}^t = \mathbf{\Lambda}$. The principal form of $\mathbf{R}$ is given as

$$\begin{aligned}\mathbf{R}_{(n\times n)} &= \mathbf{D\Lambda D}^t_{(n\times n)}\\ &= \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m}\\ d_{21} & d_{22} & \cdots & d_{2m}\\ \vdots & \vdots & \cdots & \vdots\\ d_{n1} & d_{n2} & \cdots & d_{nm}\end{bmatrix}\\ &\quad\times \begin{bmatrix} \lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & \lambda_m\end{bmatrix} \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m}\\ d_{21} & d_{22} & \cdots & d_{2m}\\ \vdots & \vdots & \cdots & \vdots\\ d_{n1} & d_{n2} & \cdots & d_{nm}\end{bmatrix},\end{aligned} \tag{1}$$

where columns of $\mathbf{D}$ and $\mathbf{D}^t$ are the eigenvectors and diagonal entries of $\mathbf{\Lambda}$ are the eigenvalues. Components whose eigenvalues exceed unity, $\lambda_j > 1$, are extracted from $\mathbf{\Lambda}$ and sorted such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 1$. The "loading" or correlation between genes and extracted components is represented by a matrix in the form

$$\mathbf{L}_{(n\times m)} = \begin{bmatrix} \sqrt{\lambda_1} d_{11} & \sqrt{\lambda_2} d_{12} & \cdots & \sqrt{\lambda_m} d_{1m} \\ \sqrt{\lambda_1} d_{21} & \sqrt{\lambda_2} d_{22} & \cdots & \sqrt{\lambda_m} d_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \sqrt{\lambda_1} d_{n1} & \sqrt{\lambda_2} d_{n2} & \cdots & \sqrt{\lambda_m} d_{nm} \end{bmatrix}, \qquad (2)$$

where rows represent genes and columns represent components, and, for example, $\sqrt{\lambda_1} d_{11}$ is the loading (correlation) between gene 1 and component 1. CLUSFAVOR algorithm proposed by Leif [41] performs PCA along with hierarchical clustering (see "Hierarchical clustering and decision tree" section) with DNA microarray expression data. CLUSFAVOR standardizes expression data and sorts and performs hierarchical and PCA of arrays and genes. Applying CLUSFAVOR, principal component method is used and component extraction and loading calculations are completed, a *varimax* orthogonal rotation of components is completed so that each gene mostly loads on a single component [42]. The result reported in [41] mixing hierarchical clustering and PCS was summarized through a colored tree, where genes that load strongly negative (less than -0.45) or strongly positive (greater than 0.45) on a single component are indicated by the use of two arbitrary colors in the column for each component whereas genes with identical color patterns in one or more columns were considered as having similar expression profiles within the selected group of genes.

Unsupervised learning based on normal mixture models

Unsupervised clustering is used to detect pattern, feature discovery, and also to match the protein sequence to the database sequences. Unsupervised learning enables pattern discovery by organizing data into clusters, using recursive partitioning methods. In the last 25 years it has been found that basing cluster analysis on a probability model can be useful both for understanding when existing methods are likely to be successful and for suggesting new methods [43, 44, 45, 46, 47, 48, 49]. One such probability model is that the population of interest consists of K different subpopulations $G_1, \ldots, G_K$ and that the density of a p-dimensional observation $\mathbf{x}$ from the kth subpopulation is $f_k(\mathbf{x}, \theta_k)$ for some unknown vector of parameters θ_k $(k = 1, \ldots, K)$. Given observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we let $\nu = (\nu_1, \ldots, \nu_n)^t$ denote the unknown identifying labels, where $\nu_i = k$ if $\mathbf{x}_i$ comes from the kth subpopulation. In the so-called classification maximum likelihood procedure, $\theta = (\theta_1, \ldots, \theta_K)$ and $\nu = (\nu_1, \ldots, \nu_n)^t$ are chosen to maximize the classification likelihood:

$$p(\theta_1, \ldots, \theta_K; \nu_1, \ldots, \nu_n | \mathbf{x}) = \prod_{i=1}^{n} f_{\nu_i}(\mathbf{x}_i | \theta_{\nu_i}). \qquad (3)$$

Normal mixture is a traditional statistical tool which has successfully been applied in gene expression [50]. For multivariate data of a continuous nature, attention has focused on the use of multivariate normal components because of their computational convenience. In this case, the data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ to be classified are viewed as coming from a mixture of probability distributions, each representing a different cluster, so the likelihood is expressed as

$$p(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K | \mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i | \theta_k), \qquad (4)$$

where π_k is the probability that an observation belongs to the kth components ($\pi_k \geq 0$; $\sum_{k=1}^{K} \pi_k = 1$).

In the theory of finite mixture, recently, methods based on this theory performed well in many cases and applications including character recognition [51], tissue segmentation [52], application to astronomical data [53, 54, 55] and enzymatic activity in the blood [56].

Once the mixture is fitted, a probabilistic clustering of the data into a certain number of clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. The likelihood ratio statistic, Bayesian information criteria (BIC), Akaike information criteria (AIC), information complexity criteria (ICOMP), and others are used to choose the number of clusters if there is any. A mixture of t-distribution may also be used instead of mixture of normals in order to provide some protection against atypical observations, which are prevalent in microarray data.

McLachlan et al [50] proposed a model-based approach to the clustering of tissue samples on a very large number of genes. They first select a subset of genes relevant for the clustering of the tissue samples by fitting mixtures of t distributions to rank the genes in order of increasing size of the likelihood ratio statistic for the test of one versus two components in the mixture model. The use of t component distributions was employed in the gene selection in order to provide some protection against atypical observations, which exit in genomics and proteomics data. In this case, the data $\mathbf{x}$ to be classified is viewed as coming from a mixture of probability distributions (4), where $f_k(\mathbf{x} | \theta_k = (\mu_k, \Sigma_k, \gamma_k))$ is a t density with location μ_k, positive definite inner product matrix Σ_k, and γ_k degrees of freedom is given by

$$\frac{\Gamma((\gamma_k + p)/2) \left| \boldsymbol{\Sigma}_k \right|^{-1/2}}{(3.14 \times \gamma_k)^{1/2} \Gamma(\gamma_k/2) \{1 + \delta(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)/\gamma_k\}^{(1/2)(\gamma_k + p)}}, \qquad (5)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k) = (\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k (\mathbf{x} - \boldsymbol{\mu}_k)$ denotes the Mahalanobis squared distance between $\mathbf{x}$ and $\boldsymbol{\mu}_k$. If $\gamma_k > 1$, $\boldsymbol{\mu}_k$ is the mean of $\mathbf{x}$ and $\gamma_k > 2$, $\gamma_k(\gamma_k - 2)^{-1}\Sigma_k$ is its covariance matrix.

McLachlan approach was demonstrated on two well-known data sets on colon and leukemia tissues. The algorithm proposed is used to select relevant genes for clustering the tissue samples into two clusters corresponding to healthy and unhealthy tissues.

Weighted voting (WV)

The weighted voting (WV) algorithm directly applies the signal-to-noise ratio to perform binary classification. For a chosen feature **x** of a test sample, it measures its distance with respect to decision boundary $b = (1/2)(\mu_1 + \mu_2)$, which is located halfway between the average expression levels of two classes, where μ_1 and μ_2 are the centers of the two clusters. If the value of this feature falls on one side of the boundary, a vote is added to the corresponding class. The vote $V(\mathbf{x}) = P(g,c)(\mathbf{x} - b)$ is weighted by the distance between the feature value and the decision boundary and the signal-to-noise ratio of this feature determined by the training set. The vote for each class is computed by summing up the weighted votes, $V(\mathbf{x})$, made by selected features for this class. In this contest, Yeang et al [57] performed multiclass classification by combining the outputs of binary classifiers. Three classifiers including weighted voting were applied over 190 samples from 14 tumor classes where a combined expression dataset was generated. Weighted Voting is a classification tool which, based on the already known clusters, proposes a rule of classification of the data set and then predicts the allocation of new samples to one of the established clusters.

k-nearest neighbors (kNN)

The *k*NN algorithm is a popular instance-based method of cluster analysis. The algorithm partitions data into a predetermined number of categories as instances are examined, according to a distance measure (eg, Euclidean). Category centroids are fixed at random positions when the model is initialized, which can affect the clustering outcome.

*k*NN is popular because of its simplicity. It is widely used in machine learning and has numerous variations [58]. Given a test sample of unknown label, it finds the *k* nearest neighbors in the training set and assigns the label of the test sample according to the labels of those neighbors. The vote from each neighbor is weighted by its rank in terms of the distance to the test sample.

Let $G_m = (g_{1m}, g_{2m}, \ldots, g_{qm})$, where g_{im} is the log expression ratio of the *i*th gene in the *m*th specimen; $m = 1, \ldots, M$ (M = number of samples in the training set). In the *k*NN method, one computes the Euclidean distance between each specimen, represented by its vector G_m, and each of the other specimens. Each specimen is classified according to the class membership of its *k*-nearest neighbors. In a study undertaken by Hamadeh et al [59], the training set comprised of RNA samples derived from livers of Sprague-Dawley rats exposed to one of 3 peroxisome proliferations. In this study, $M = 27$, $q = 30$, and $k = 3$. A set of q ($q = 30$) genes was considered discriminative when at least 25 out of 27 specimens were correctly classified. A total of 10,000 such subsets of genes were obtained. Genes were then rank-ordered according to how many times they were selected into these subsets. The top 100 genes were subsequently used for prediction purposes.

*k*NN can also be used for recovering missing values in DNA microarray. In fact, hundreds of genes can be observed in one particular experiment. Arrays are printed with approximately 1 kilobase of DNA, corresponding to the coding region of a particular gene, per spot. Labelling of cDNA is done to determine where hybridization occurs. Hybridization is viewed either by fluorescence or radioactive intensity. One drawback of these techniques is the scanning of hybridization intensities. A certain threshold value must be met in order for a value to be returned as a valid measurement. If a value is below this threshold, it is returned as missing data. This missing data disrupts the analysis of the experiment. For instance, if a gene is printed in a duplicate, over a series of arrays, and one spot on one array is below the threshold, the gene is disregarded across all arrays. The loss of this gene expression data is costly because no experimental conclusions can be made from the loss of expression of this gene over all arrays [60].

Artificial neural network (ANN)

Unsupervised neural networks provide a more robust and accurate approach to the clustering of large amounts of noisy data. Neural networks have a series of properties that make them suitable for the analysis of gene expression and proteins patterns. They can deal with real-world data sets containing noisy, ill-defined items with irrelevant variables and outliers, and whose statistical distribution does not need to be parametric. Multilayer perceptrons [61] provide a nonlinear mapping where the real-valued input **x** is transformed and mapped to get a real-valued output **y**:

$$\mathbf{x} \longrightarrow \mathbf{W} \times \mathbf{x} \longrightarrow \mathbf{h} \longrightarrow \mathbf{y}, \tag{6}$$

where **W** is the weight matrix, called first layer, **h** is a nonlinear transformation, **y** is a finished node. The following is an example of a two-layer neural network:

$$\begin{aligned}
\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} &\longrightarrow \mathbf{W} \times \mathbf{x} = \begin{pmatrix} \sum_{i=1}^{4} \alpha_i x_i \\ \sum_{i=1}^{4} \beta_i x_i \end{pmatrix} \\
&= \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \longrightarrow \begin{pmatrix} h(\alpha_1) = \dfrac{1}{1+e^{-\alpha_1}} \\ h(\alpha_2) = \dfrac{1}{1+e^{-\alpha_2}} \end{pmatrix}, \\
y &= \sum_{i=1}^{2} w_i h_i
\end{aligned} \tag{7}$$

if $0 < y < 1$, then we have a classification case with two groups. Technically, classification, for example, is achieved

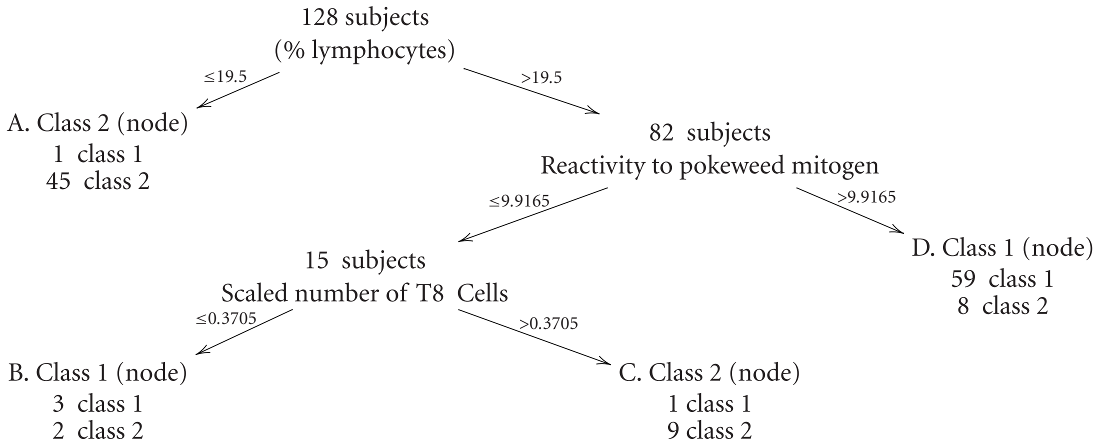

Figure 1. An example of neural network black box: a four-dimensional data input **x** is first transformed by **W**, then by h in order to give a grouping variable **y** as an output.

by comparing $y = h(\mathbf{x})$ with a threshold, we suppose here 0 for simplicity, if $h(\mathbf{x}) > 0$, observation **x** belongs to the cluster 1, if $h(\mathbf{x}) < 0$, then **x** belongs to cluster 2. The weights **W** are estimated by examining the training points sequentially.

ANN has been applied to a number of diverse areas for the identification of "biologically relevant" molecules, including pyrolysis mass spectrometry [62] and genomics microarraying of tumor tissue [63]. Ball et al [64] utilized a multilayer perceptron with a back propagation algorithm for the analysis of SELDI mass spectrometry data. This type of ANN is a powerful tool for the analysis of complex data [65]. Wei et al [66] used the same algorithm for data containing a high background noise. ANN can be used to identify the influence of many interacting factors [67] that makes it highly suitable for the study of first-generation SELDI-derived data. It can be used for the classification of human tumors and rapid identification of potential biomarkers [64]. ANN can produce generalized models with a greater accuracy than conventional statistical techniques in medical diagnostics [68, 69] without relying on predetermined relationships as in other modeling techniques. Usually, the data needs to be trained when using ANN to predict tumor grade; also the choice of the number of layers has to be proposed. Currently, ANN does not propose criteria for choosing the number of layers which should be investigator-proposed. A criteria has to be developed for the ANN to choose the adequate number of layers.

For the probabilistic modeling, usually the normality is assumed, whereas in the ANN the data is distribution-free, which makes the ANN a powerful tool for data analysis [70].

Hierarchical clustering and decision tree

The basic idea of the tree is to partition the input space recursively into two halves and approximate the function in each half by the average output value of the samples it contains [71]. Each bifurcation is parallel to one of the axes and can be expressed as an inequality involving the input components (eg, $\mathbf{x}_k > a$). The input space is divided into hypertangles organized into a binary tree where each branch is determined by the dimension (k) and boundary (a) which together minimize the residual error between model and data.

Example

In a study undertaken by Robert Dillman at the University of California, San Diego Cancer Center [72], 21 continuous laboratory variables related to immunocompetence, age, sex, and smoking habits in an attempt to distinguish patient with cancer. Prior probabilities are chosen to be equal: $\pi(1) = \pi(2) = 0.5$, and $C(1|2)$, the cost of misclassification, was calculated. The tree in Figure 1 summarizes the classification of 128 observations into two classes: supposedly healthy and unhealthy.

Currently, hierarchical clustering is the most popular technique employed for microarray data analysis and gene expression [73]. Hierarchical methods are based on building a distance matrix summarizing all the pairwise similarities between expression profiles, and then generating cluster trees (also called dendrograms) from this matrix. Genes which appear to be coexpressed at various time points are positioned close to one another in the tree whose branches lengths represent the degree of similarity between expression profiles.

Decision trees [74] were used to classify proteins as either soluble or insoluble, based on features of their amino acid sequences. Useful rules relating these features with protein solubility were then determined by tracing the paths through the decision trees. Protein solubility strongly influences whether a given protein is a feasible target for structure determination, so the ability to predict this property can be a valuable asset in the optimization of high-throughput projects. These techniques have already been applied to the study of gene expression patterns [73]. Neverthless, classical hierarchical clustering presents drawbacks when dealing with data containing a nonnegligible amount of noise. Hierarchical clustering suffers from a lack of robustness and solutions may not be unique and dependent on the data order. Also, the deterministic nature of hierarchical clustering and the impossibility of re-evaluating the results in the light of the complete data can cause some clusters of patterns to be based on local decisions rather than on the global picture.

Self-organizing mapping (SOM)

The self-organizing feature map (SOM) [75] consists of a neural network whose nodes move in relation to category membership. As with k-means, a distance measure is computed to determine the closest category centroid. Unlike k-means, this category is represented by a node with an associated weight vector. The weight vector of the matching node, along with those of neighboring nodes, is updated to more closely match the input vector. As data points are clustered and category centroids are updated, the positions of neighboring nodes move in relation to them. The number of network nodes which constitute this neighborhood typically decreases over time. The input space is defined by the experimental input data, whereas the output space consists of a set of nodes arranged according to certain topologies, usually two-dimensional grids. The application of the algorithm maps the input space onto the smaller output space, producing a reduction in the complexity of the analyzed data set [76, 77]. Like PCA, the SOM is capable of reducing high-dimensional data into a 1- or 2-dimensional representation. The algorithm produces a topology-preserving map, conserving the relationships among data points. Thus, although either method may be used to effectively partition the input space into clusters of similar data points, the SOM can also indicate relationships between clusters.

SOM is reasonably fast and can be easily scaled to large data sets. They can also provide a partial structure of clusters that facilitate the interpretation of the results. SOM structure, unlike the case of hierarchical cluster, is a two-dimensional grid usually of hexagonal or rectangular geometry, having a number of nodes fixed from the beginning. The nodes of the network are initially random patterns. During the training process, that implies slight changes in the nodes after repeated comparison with the data set, the node changes in a way that captures the distribution of variability of the data set. In this way, similar gene, peak, protein profile patterns map close together in the network and, as far as possible from the different patterns.

A combination of SOM and decision tree was proposed by Herrero et al [78]. The description of the algorithm is given as follows: given the patterns of expression that has to be classified, if two genes are described by their expression patterns as $g_1(e_{11}, e_{12}, \ldots, e_{1n})$ and $g_2(e_{21}, e_{22}, \ldots, e_{2n})$ and their distance $d_{1,2} = \sqrt{\sum(e_{1i} - e_{2i})^2}$, the initial system of the SOM is composed of two external elements, connected by an internal element. Each cell is a vector with the same size as the gene profiles. The entries of the two cells and the node are initialized. The network is trained only through their terminal neurons or cells. The algorithm proceeds by expanding the output topology starting from the cell having the most heterogeneous population of associated input gene profiles. Two new descendents are generated from this heterogeneous cell that changes its state from cell to node. The series of operations performed until a cell generates two descendents is called a cycle. During a cycle, cells and nodes are repeatedly adapted by the input gene profiles. This process of successive cycles of generation of descendant cells can last until each cell has one single input gene profile assigned (or several, identical profiles), producing a complete classification of all the gene profiles. Alternatively, the expansion can be stopped at the desired level of heterogeneity in the cells, producing in this way a classification of profiles at a higher hierarchical level.

Kanaya et al [79] use SOM to efficiently and comprehensively analyze codon usage in approximately 60,000 genes from 29 bacterial species simultaneously. They showed that SOM is an efficient tool for characterizing horizontally transferred genes and predicting the donor/acceptor relationship with respect to the transferred genes. They examined codon usage heterogeneity in the *E coli O* 157 genome, which contains the unique segments including O-islands [81] that are absent in *E coli K* 12.

Support vector machine (SVM)

SVM originally introduced by Vapnik and coworkers [82, 83] is a supervised machine learning technique. SVMs are a relatively new type of learning algorithms [84, 85] successively extended by a number of researchers. Their remarkably robust performance with respect to sparse and noisy data is making them the system of choice in a number of applications from text categorization to protein function prediction. SVM has been shown to perform well in multiple area of biological analysis including evaluating microarray expression data [86], detecting remote protein homologies, and recognizing translation initiation sites [87, 88, 89]. When used for classification, they separate a given set of binary-labeled training data with a hyperplane that is maximally distant from them known as "the maximal margin hyperplane." For cases in which no linear separation is possible, they can work in combination with the technique of "kernels" that automatically realizes a nonlinear mapping to a feature space.

The SVM learning algorithm finds a hyperplane $(\mathbf{w}, \mathbf{b})$ such that the margin γ is maximized.The margin γ is defined as a function of distance between the input $\mathbf{x}$, labeled by the random variable $\mathbf{y}$, to be classified and the decision boundary $(\langle \mathbf{w}, \phi(\mathbf{x})\rangle - \mathbf{b})$:

$$\gamma = \min_{\mathbf{x}} \operatorname{sign}\{\langle \mathbf{w}, \phi(\mathbf{x})\rangle - \mathbf{b}\}, \tag{8}$$

where ϕ is a mapping function from the input space to the feature space.

The decision function to classify a new input x is

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i \mathbf{y}_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x})\rangle - \mathbf{b}\right). \tag{9}$$

When the data is not linearly separable, one can use more general functions that provide nonlinear decision boundaries, like polynomial kernels

$$K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle = (\langle \mathbf{x}_i, \mathbf{x}_j\rangle + 1)^p \tag{10}$$

or Gaussian kernels $K_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|/\sigma^2}$, where p and σ are kernel parameters.

To apply the SVM for gene classification, a set of examples was assembled containing genes of known function, along with their corresponding microarray expression profiles. The SVM was then used to predict the functions of uncharacterized yeast open reading frames (ORFs) based on the expression-to-function mapping established during training [86]. Supervised learning techniques appear to be ideal for this type of functional classification of microarray targets, where sets of positive and negative examples can be compiled from genomic sequence annotations.

Boolean network

The basis for the Boolean networks was introduced by Turing and von Neumann in the form of automata theory [90, 91]. A Boolean network is a system of n interconnected binary elements; any element in the system can be connected to a series I of other k elements, where k (and hence I) can vary. For each individual element, there is a logical or Boolean rule B which computes its value based on the values of elements connected with one. The state of the system S is defined by the pattern of states (on/off or 0/1) of all elements. All elements are updated synchronously, moving the system into its next state, and each state can have only one resultant state. The total system space is defined as all possible N combinations of the values of the n elements in S.

One of the important types of information underlying the expression profile data is the regulatory networks among genes, which is called also "genetic network." Modeling with the Boolean network [92, 93, 94, 95] has been investigated for inferences of the genetic networks. Tavazoie et al [96] proposed an approach that combines cluster analysis with sequence motif detection to determine the genetic network architecture. Recently, an approach to infer the genetic networks with Bayesian networks was proposed [97] but still a little has been done in this area using Boolean network.

Combination of cluster analysis and a graphical Gaussian modeling (GGM)

GGM is an algorithm that was proposed by Toh and Horimoto [98] to cluster expression profile data. GGM is a multivariate analysis to infer or test a statistical model for the relationship among a plural of variables, where a partial correlation coefficient, instead of a correlation coefficient, is used as a measure to select the first type of interaction [99, 100]. In GGM, the statistical model for the relationship among the variables is represented as a graph, called the "independence graph," where the nodes correspond to the variables under consideration and the edges correspond to the first type of interaction between variables. More specifically, an edge in the independence graph indicates a pair of variables that are conditionally dependent. GGM was applied for the expression profile data of 2467 *Saccharomyces cerevisiae* genes measured under 79 different conditions [73]. The 2467 genes were classified into 34 clusters by a cluster analysis, as a preprocessing for GGM. Then the expression levels of the genes in each cluster were averaged for each condition. The averaged expression profile data of 34 clusters were subjected to GGM and a partial correlation coefficient matrix was obtained as a model of the genetic network of the *S cerevisiae.*

Other probabilistic and clustering methods and applications

To try to make a sense to microarray data distributions, Hoyle et al [101] proposed a comparison of the entire distribution of spot intensities between experiments and between organisms. The novelty of this study is by showing that there is a close agreement with Benford's law and Zipf's law [102, 103] which is a combination of lognormal distribution of large majority of the spot intensity values and the Zipf's law for the tail.

In addition to the clustering methods that we have described, there exist numerous other methods. Bensmail and Celeux [104] used model-based cluster analysis to cluster 242 cases of various grades of neoplasia which were collected and diagnosed in a subsequently taken biopsy [105]. There were 50 cases with mild displasia, 50 cases with moderate displasia, 50 cases with severe displasia, 50 cases with carcinoma in situ, and 42 cases with invasive carcinoma. Eleven measurements were used in this study, 7 are ordinal and 4 are numerical. Using eigenvalue decomposition regularized discriminant analysis algorithm (EDRDA), 14 models were investigated and their performance was measured by their error rate of misclassification with cross-validation. Each model describes a specific orientation, shape, and volume of the cluster defined by the spectral decomposition of the covariance matrix Σ_k related to each cluster:

$$\Sigma_k = \lambda_k D_k A_k D_k^t, \tag{11}$$

Table 1. Summary of the 14 models presented in Bensmail and Celeux [104].

Model 1 = $[\lambda DAD^t]$	Model 2 = $[\lambda_k DAD^t]$	Model 3 = $[\lambda DA_kD^t]$	Model 4 = $[\lambda_k DA_kD^t]$
Model 5 = $[\lambda D_kAD_k^t]$	Model 6 = $[\lambda_k D_kAD_k^t]$	Model 7 = $[\lambda D_kA_kD_k^t]$	Model 8 = $[\lambda_k D_kA_kD_k^t]$
Model 9 = $[\lambda I]$	Model 10 = $[\lambda_k I]$	Model 11 = $[\lambda B]$	Model 12 = $[\lambda_k B]$
Model 13 = $[\lambda B_k]$	Model 14 = $[\lambda_k B_k]$		

Table 2. Summary of the properties of the most commonly applied algorithms for data analysis.

	Time/space	Strengths	Weaknesses
PCA	$(p(p+1)/2)$ p: no. of variables	Dimension reduction	Circular shape
Unsupervised learning normal mixture	$(kp^2n)/\ O(kn)$ p: no. of variables k: no. of clusters	Clustering and prediction	Normality assumption
Weighted voting	(kp) p: no. of variables k: no. of clusters	Tailored weights Weights flexibility	Binary classification
kNN	(tkn) k: no. of clusters n: no. of observations t: no. of iterations	Image processing Handling missing data	Known mean Known number of classes
ANN	$O(n)$ n: no. of observations	Nonlinear/Noisy data	Black box behavior
Hierarchical/tree	$O(n^2)$ n: no. of observations	Readability of results	Numerical data only No scaling of data
SOM	$O(n)$ n: no. of observations	Topology preserving Computationally tractable Handling high dimension	Trained on normal data No reliability
SVM	$O(n^2)$ n: no. of observations	Easy training Handling high-dimensional data	Need to a kernel function
Boolean network	$O(n(d))$ n: no. nodes d: max(indegree)	Defining relationships	No handling of missing data Trained on large data
GGM	$O(kp^2)$ k: no. of clusters p: no. of variables	Probabilistic model Graphical model	Conditional probability
Model-based	$O(kp^2n)$ k: no. of clusters n: no. of observations p: no. of variables	Geometry of the clusters	Normality

where $\lambda_k = |\Sigma_k|^{1/p}$ describes the volume of the cluster G_k, D_k, the eigenvectors matrix, describes the orientation of the cluster G_k, and A_k, the eigenvalues matrix, describes the shape of the cluster G_k. Table 1 summarizes the fourteen models.

This methodology seems very promising since it took in consideration the characteristics of the clusters (shape, volume, and orientation) and then proposed a flexible way of discriminating the data by proposing a panoply of rules varying from the simple one (linear discriminant rule) to the complex one (quadratic discriminant rule). This methodology can easily be applied to discriminate/classify peaks of protein profiles when they are appropriately transformed. Since EDRDA is based on the assumption that the data is distributed according to a mixture of Gaussian distributions, some extent to which different transformations of gene expression or protein profiles sets satisfying the normality assumption may be explored. Three commonly used transformations can be applied: logarithm, square root, and standardization (wherein the raw expression levels for each gene [protein profile] are transformed by substracting their mean and dividing by their standard deviation) [106]. Other more interesting transformations may be investigated including kernel smoother.

The summary of the above-described methods for clustering, classification, and prediction of gene expression and protein profiles sets is presented in Table 2. We present the algorithms, their performance, their strengths, and weaknesses. Over all, some methods are efficient for some applications such as imputing data but performs less in clustering. Probabilistic methods such as model-based methods and mixture models are interesting to look at after transforming the data sets because they are a natural fit to cluster data sets with underlying distribution. Nonprobabilistic methods such as the Neural network and the Kohonen mapping may be interesting when the data contains an important amount of noise.

CONCLUSION

The postgenomic era holds phenomenal promise for identifying the mechanistic bases of organismal development, metabolic processes, and disease, and we can confidently predict that bioinformatics research will have a dramatic impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and drug discovery.

Software packages and bioinformatic tools have been and are being developed to analyze 2D gel protein patterns. These software applications possess user-friendly interfaces that are incorporated with tools for linearization and merging of scanned images. The tools also help in segmentation and detection of protein spots on the images, matching, and editing [107]. Additional features include pattern recognition capabilities and the ability to perform multivariate statistics. The handling and analysis of the type of data to be collected in proteomic investigations represent an emerging field [Bensmail H, Hespen J. Semmes OJ, and Haudi A. Fast Fourier transform for Bayesian clustering of Proteomics data (unpublished data).]. New techniques and new collaborations between computer scientists, biostatisticians, and biologists are called for. There is a need to develop and integrate database repositories for the various sources of data being collected, to develop tools for transforming raw primary data into forms suitable for public dissemination or formal data analysis, to obtain and develop user interfaces to store, retrieve, and visualize data from databases and to develop efficient and valid methods of data analysis.

REFERENCES

[1] O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. 1975;250 (10):4007–4021.

[2] Merril CR, Switzer RC, Van Keuren ML. Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. *Proc Natl Acad Sci USA*. 1979;76(9):4335–4339.

[3] Patton WF. Making blind robots see: the synergy between fluorescent dyes and imaging devices in automated proteomics. *Biotechniques*. 2000;28(5):944–957.

[4] Steinberg TH, Jones LJ, Haugland RP, Singer VL. SYPRO orange and SYPRO red protein gel stains: one-step fluorescent staining of denaturing gels for detection of nanogram levels of protein. *Anal Biochem*. 1996;239(2):223–237.

[5] Chambers G, Lawrie L, Cash P, Murray GI. Proteomics: a new approach to the study of disease. *J Pathol*. 2000;192(3):280–288.

[6] Bergman AC, Benjamin T, Alaiya A, et al. Identification of gel-separated tumor marker proteins by mass spectrometry. *Electrophoresis*. 2000; 21(3):679–686.

[7] Chakravarti DN, Chakravarti B, Moutsatsos I. Informatic tools for proteome profiling. *Biotechniques*. 2002;32(Suppl):4–15.

[8] Lopez MF, Kristal BS, Chernokalskaya E, et al. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis*. 2000;21(16):3427–3440.

[9] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*. 1988;60(20):2299–2301.

[10] Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem*. 1991;63 (24):1193A–1203A.

[11] Andersen JS, Mann M. Functional genomics by mass spectrometry. *FEBS Lett*. 2000;480(1):25–31.

[12] Krutchinsky AN, Zhang W, Chait BT. Rapidly switchable matrix-assisted laser desorption/ionization and electrospray quadrupole-time-of-flight mass spectrometry for protein identification. *J Am Soc Mass Spectrom*. 2000;11(6):493–504.

[13] Shevchenko A, Loboda A, Shevchenko A, Ens W, Standing KG. MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. *Anal Chem*. 2000;72(9):2132–2141.

[14] Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*. 2000;21(6):1164–1177.

[15] Wright Jr GL, Cazares LH, Leung SM, et al. Proteinchip® surface enhanced laser desorption/

ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis*. 1999;2(5-6):264–276.

[16] Vlahou A, Schellhammer PF, Mendrinos S, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am J Pathol*. 2001;158(4):1491–1502.

[17] 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.

[18] Geysen HM, Meloen RH, Barteling SJ. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci USA*. 1984;81(13):3998–4002.

[19] De Wildt RM, Mundy CR, Gorick BD, Tomlinson IM. Antibody arrays for high-throughput screening of antibody-antigen interactions. *Nat Biotechnol*. 2000;18(9):989–994.

[20] Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A. Protein microchips: use for immunoassay and enzymatic reactions. *Anal Biochem*. 2000;278(2):123–131.

[21] Haab BB, Dunham MJ, Brown PO. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol*. 2001;2(2):1–13.

[22] Cahill DJ. Protein and antibody arrays and their medical applications. *J Immunol Methods*. 2001; 250(1-2):81–91.

[23] Kononen J, Bubendorf L, Kallioniemi A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*. 1998; 4(7):844–847.

[24] Hoos A, Urist MJ, Stojadinovic A, et al. Validation of tissue microarrays for immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. *Am J Pathol*. 2001;158(4):1245–1251.

[25] Camp RL, Carette LA, Rimm DL. Validation of tissue microarray technology in breast cancer. *Lab Invest*. 2000;80:1943-1949.

[26] Mucci NR, Akdas G, Manely S, Rubin MA. Neuroendocrine expression in metastatic prostate cancer: evaluation of high throughput tissue microarrays to detect heterogeneous protein expression. *Hum Pathol*. 2000;31(4):406–414.

[27] Banks RE, Dunn MJ, Hochstrasser DF, et al. Proteomics: new perspectives, new biomedical opportunities. *Lancet*. 2000;356(92430):1749–1756.

[28] Anderson NL, Matheson AD, Steiner S. Proteomics: applications in basic and applied biology. *Curr Opin Biotechnol*. 2000;11(4):408–412.

[29] Vercoutter-Edouart AS, Lemoine J, Le Bourhis X, et al. Proteomic analysis reveals that 14-3-3 sigma is down-regulated in human breast cancer cells. *Cancer Res*. 2001;61(1):76–80.

[30] Ferguson AT, Evron E, Umbricht CB, et al. High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer. *Proc Natl Acad Sci USA*. 2000;97(11):6049–6054.

[31] Chaurand P, Stoeckli M, Caprioli RM. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal Chem*. 1999;71(23):5263–5270.

[32] Stoeckli M, Chaurand P, Hallahan DE, Caprioli RM. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat Med*. 2001;7(4):493–496.

[33] Hutchens TW, Yip TT. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun Mass spectrum*. 1993;7:576–580.

[34] Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*. 2002;48(8):1296–1304.

[35] Paweletz CP, Gillespie JW, Ornstein DK, et al. Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Research*. 2000;49:34–42.

[36] Neubauer G, King A, Rappsilber J, et al. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet*. 1998;20(1):46–50.

[37] Kuster B, Mortensen P, Mann M. Identifying proteins in genome databases using mass spectrometry. In *Proceedings of the 47th ASMS Conference of Mass Spectrometry and Allied Topics*. Dallas, Tex: American Society for Mass Spectrometry; 1999: 1897–1898.

[38] Baldi P, Brunak S. *Bioinformatics: the Machine Learning Approach*. Cambridge, Mass: MIT Press; 1998.

[39] Chapman PF, Falinska AM, Knevett SG, Ramsay MF. Genes, models and Alzheimer's disease. *Trends Genet*. 2001;17(5):254–261.

[40] Keegan LP, Gallo A, O'Connell MA. Development. Survival is impossible without an editor. *Science*. 2000;290(54970)1707–1709.

[41] Peterson LE. CLUSFAVOR 5.0: hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. *Genome Biology*. 2002;3(7)1–8.

[42] Kaiser HF. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. 1958;23:187–200.

[43] Binder DA. Bayesian cluster analysis. *Biometrika*. 1978;65:31–38.

[44] Hartigan JA. *Clustering Algorithms*. New York, NY: John Wiley & Sons; 1975.

[45] Menzefricke U. Bayesian clustering of data sets.

Communications in Statistics. 1981;A10:65–77.

[46] Symons MJ. Clustering criteria and multivariate normal mixtures. *Biometrics*. 1981;37:35–43.

[47] McLachlan GJ. The classification and mixture maximum likelihood approaches to cluster analysis. In: Krishnaiah PR, Kanal LN, eds. *Handbook of Statistics*. vol.2 Amsterdam, Holland: North-Holland Publishing; 1982:199–208.

[48] McLachlan GJ, Basford KE. *Mixture Models: Inference and Applications to Clustering*. New York, NY: Marcel Dekker; 1988.

[49] Bock HH. Probability models in partitional cluster analysis. *Computational Statistics and Data Analysis*. 1996;23:5–28.

[50] McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*. 2002;18(3):413–422.

[51] Murtagh F, Raftery AE. Fitting straight lines to point patterns. *Pattern Recognition*. 1984;17:479–483.

[52] Banfield JD and Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics*. 1993;49:803–821.

[53] Bensmail H, Celeux G, Raftery AE, Robert C. Inference in model-based cluster analysis. *Computing and Statistics*. 1997;1(10):1–10.

[54] Roeder K, Wasserman L. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*. 1997;92: 894–902.

[55] Mukerjee ED, Feigelson GJ, Babu F, Murtagh C, Fraley C, Raftery AE. Three types of gamma ray bursts. *Astrophysical Journal*. 1998;50:314–327.

[56] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components, with discussion. *Journal of the Royal Statistical Society, B*. 1997;59(4):731–792.

[57] Yeang CH, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types. *Bioinformatics*. 2001;17(suppl 1):S316–S322.

[58] Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York, NY: John Wiley & Sons; 2001.

[59] Hamadeh HK, Bushel PR, Jayadev S, et al. Prediction of compound signature using high density gene expression profiling. *Toxicol Sci*. 2002; 67(2):232–240.

[60] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(16):520–525.

[61] Minsky M, Papert S. *Perceptrons: an Introduction to Computational Geometry*. Cambridge, Mass: MIT Press; 1969.

[62] Goodacre R, Kell DB. Pyrolysis mass spectrometry and its applications in biotechnology. *Curr Opin Biotechnol*. 1996;7(1):20–28.

[63] Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. 2001;7(6):673–679.

[64] Ball G, Mian S, Holding F, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*. 2002;18(3):395–404.

[65] De Silva CJS, Choong PL, Attikiouzel Y. Artificial neural networks and breast cancer prognosis. *Australian Computer Journal*. 1994;26(3):78–81.

[66] Wei JT, Zhang Z, Barnhill SD, Madyastha KR, Zhang H, Oesterling JE. Understanding artificial neural networks and exploring their potential applications for the practicing urologist. *Urology*. 1998;52(2):161–172.

[67] Kothari SC, Heekuck OH. Neural networks for pattern recognition. *Advances in Computers*. 1993;37:119–166.

[68] Tafeit E, Reibnegger G. Artificial neural networks in laboratory medicine and medical outcome prediction. *Clin Chem Lab Med*. 1999;37(9):845–853.

[69] Reckwitz T, Potter SR, Snow PB, Zhang Z, Veltri RW, Partin AW. Artificial neural networks in urology: Update 2000. *Prostate Cancer Prostatic Dis*. 1999;2(5-6):222–226.

[70] Rumelhart DE, McCletland JL. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, Mass: MIT Press;1:1986.

[71] Breiman L, Friedman JH, Olshen JA, Stone CJ. *Classification and Regression Trees*. Belmont, Calif: Wadsworth;1984.

[72] Dillman RO, Beauregard JC, Zavanelli MI, Halliburton BL, Wormsley S, Royston I. In vivo immune restoration in advanced cancer patients after administration of thymosin fraction 5 or thymosin alpha 1. *J Biol Response Mod*. 1983;2(2):139–149.

[73] 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.

[74] Quinlan JR. *C4.5: Programs for Machine Learning. Machine Learning*. San Mateo, Calif: Morgan Kaufmann; 1993.

[75] Kohonen T. The self-organizing map. *Proceedings of the IEEE*. 1990;78:1464–1480.

[76] 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.

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

[78] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*. 2001;17(2):126–136.

[79] Kanaya S, Kinouchi M, Abe T, et al. Analysis of

codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E coli* O157 genome. *Gene*. 2001;276(1-2):89–99.

[80] Boser BE, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th ACM Workshop on Computational Learning Theory*. New York, NY: ACM Press; 1992: 144–152.

[81] Perna NT, Plunkett III G, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*. 2001;409(6819):529–533.

[82] Vapnik V. *Statistical Learning Theory*. New York, NY: John Wiley&Sons; 1998.

[83] Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press; 2000.

[84] Shawe-Taylor J, Cristianin N. Further results on the margin distribution. In: *Proc. 12th Annual Conf. on Computational Learning Theory*. New York, NY: ACM Press; 1999.

[85] Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA*. 2000;97(1):262–267.

[86] Jaakkola T, Diekhans M, Haussler D. Using the Fisher Kernel method to detect remote protein homologies. In: *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, Calif: AAAI Press; 1999: 149–158.

[87] Zien A, Rätsch G, Mika S, Scholkopf B, Lengauer T, Muller KR. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*. 2000;16(9):799–807.

[88] Mukherjee S, Tamayo P, Mesirov JP, Slonim D, Verri A, Poggio T. Support vector machine classification of microarray data. Tech. Rep. 182/AI Memo. Cambridge, Mass: CBCL;1999.

[89] Mukherjee S, Tamayo P, Mesirov JP, Slonim D, Verri A, Poggio T. Support vector machine classification of microarray data. Tech. Rep. 1677. Cambridge, Mass: MIT; 1999.

[90] Turing A. Turing machine. *Proc London Math Soc*. 1936;242:230–265.

[91] Von Neumann J. *Theory of Self-Reproducing Automata*. Burks AW. ed. Champaign, Ill: University of Illinois Press; 1966.

[92] Somogyi R, Sniegoski CA. Modeling the complexity of genetic networks: understanding multigene and pleitropic regulation. *Complexity*. 1996;1:45–63.

[93] Chen T, He HL, Church GM. Modeling gene expression with differential equations. *Proc. Pac. Symposium on Biocomputing*. 1999;4:29–40.

[94] D'haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. *Proc. Pac. Symposium on Biocomputing*. 1999;4:41–52.

[95] Akutsu T, Miyano S, Kuhara S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol*. 2000;7(3-4):331–343.

[96] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet*. 1999;22(3):281–285.

[97] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3-4):601–620.

[98] Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*. 2002;18(2):287–297.

[99] Whittaker J. *Graphical Models in Applied Multivariate Statistics*. New York, NY: John Wiley & Sons; 1990.

[100] Edwards D. *Introduction to Graphical Modelling*. New York, NY: Springer-Verlag; 1995.

[101] Hoyle DC, Rattray M, Jupp R, Brass A. Making sense of microarray data distributions. *Bioinformatics*. 2002;18(4):576–584.

[102] Benford F. The law of anomalous numbers. *Proc. Amer. Phil. Soc*. 1938;78:551–572.

[103] Zipf GK. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass: Addison-Wesley; 1949.

[104] Bensmail H, Celeux G. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American statistical Association*. 1996;91: 1743–1748.

[105] Meulman JJ, Zeppa P, Boon ME, Rietveld WJ. Prediction of various grades of cervical neoplasia on plastic-embedded cytobrush samples. Discriminant analysis with qualitative and quantitative predictors. *Anal Quant Cytol Histol*. 1992;14(1):60–72.

[106] 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.

[107] Ohler U, Harbeck S, Niemann H, Noth E, Reese MG. Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*. 1999;5(5):362–369.

* Corresponding author.
E-mail: haoudia@evms.edu
Fax: +1 757 624 2255; Tel: +1 757 446 5682