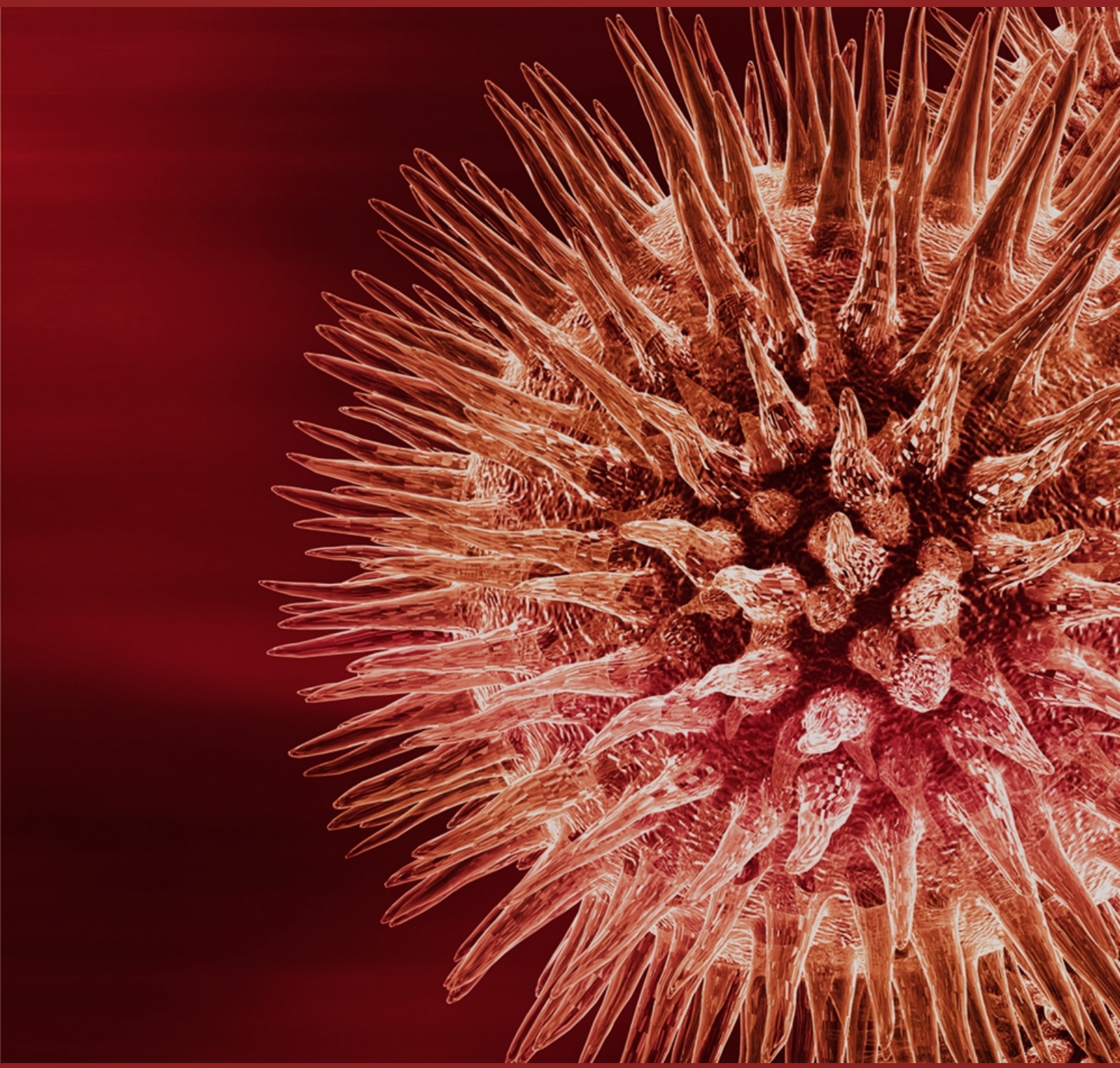
Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part I

Guest Editors: George L. Wright Jr and O. John Semmes



Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part I

Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part I

Guest Editors: George L. Wright Jr and O. John Semmes

Contents

Journal of Biomedicine and Biotechnology • 2003:4 (2003) 215–216 • PII. S1110724303002997 • http://jbb.hindawi.com

EDITORIAL

Proteomics in Health and Disease

George L. Wright Jr* and O. John Semmes†

Department of Microbiology and Molecular Cell Biology, Virginia Prostate Center, Eastern Virginia Medical School, Norfolk, VA 23501, USA

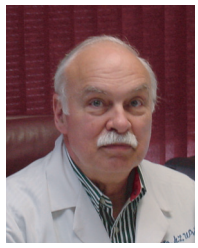
One of the major goals of the postgenomic era is understanding the structures, interactions, and functions of all cell proteins. This becomes a daunting task considering the estimation that there are between 100 000 and 200 000 individual proteins resulting from alternative splicing of the 30 000 genes encoded by the human genome. Since the cellular proteome is a dynamic profile, subject to change in response to various signals through posttranslational modification, translocation, and protein-protein and protein-nucleic acid interactions, the task becomes even more complex looming to a million or more modification events. Proteomics encompasses the study of expressed proteins, including identification and elucidation of the structure-function interrelationships which define healthy and disease conditions. Information at the level of the proteome is critical to understanding the function of cellular phenotype and its role in health and disease. Since posttranslational events and, indeed, an accurate assessment of protein expression levels cannot always be predicted by mRNA analysis, proteomics, used in concert with genomics, can provide a holistic understanding of the biology underlying the disease process. The challenge in deciphering the proteome is the development and integration of analytical instrumentation combined with bioinformatics that provide rapid, high-throughput, sensitive, and reproducible tools.

This issue of the Journal of Biomedicine and Biotechnology presents the first of a two-part series consisting of ten papers that describe both technical and bioinformatic advances to define the cell proteome towards a better understanding of health and disease. The current issue consists of the first five articles beginning with papers by Bensmail and Haoudi, and Pruess and Apweiler that describe bioinformatics approaches for defining the cancer cell proteome and for in silico proteomic analyses. Because of the high dimensionality of the data generated by proteomic methodologies, such as protein microarrays and mass spectral analyses, more efficient and accurate bioinformatics tools are required to mine and analyze the data. Major advances in mass spectrometry have resulted in rapid, high-throughput technologies for protein biomarker discovery, protein identification, disease analyses, and identification of posttranslational modifications. One advance, SELDI ProteinChip mass spectrometry, is the subject of the next two papers that describe its use for biomarker discovery and its potential as a platform for development of multimarker clinical assays. The first paper by Reddy and Dalmasso presents a review describing the use of SELDI for biomarker discovery, drug discovery, protein identification, and for development of multiplex clinical assays, citing examples for cancer, neurological disorders, and infectious diseases. Feng and associates then describe an automated peak identification and calibration procedure for more precise mass analyses when attempting to differentiate disease from nondisease protein patterns. The last paper of this issue by Qoronfleh and associates describes a method for the isolation of membrane proteins for proteomic analysis.

The next issue (volume 2003, issue 5) presents the remaining five papers. This issue begins with a review by Xu and Lam on protein and chemical microarray approaches being utilized for proteomic studies. Then Flower and colleagues describe bioinformatics approaches for defining the immunome for discovery of novel vaccines. This is followed by a paper by Qoronfleh and colleagues who describe improved methods for detecting protein: protein interactions. Piccoli's research team then report a method for optimizing the rolling circle application technology for generating a sensitive high-throughput multiplex protein microarray for analysis of protein expression and molecular diagnosis. The final paper in this issue is by Vlahou and associates who describe the use of SELDI protein profiling coupled with a commercial decision tree learning algorithm for biomarker discovery and diagnosis of ovarian cancer.

The content of this special issue, although broad and addressing several key issues in proteomics research, still leaves many issues to be covered, especially functional and structural proteomics, in this fast evolving field of research. We anticipate addressing other new discoveries and applications in the proteomics field in future issues of the Journal of Biomedicine and Biotechnology.

George L. Wright Jr
O. John Semmes



George L. Wright Jr received the PhD degree from Michigan State University in 1966. His PhD research focused on deciphering the Mycobacterium proteome and detecting changes in the serum proteome of cows infected with tuberculosis using one- and two-dimensional electrophoresis. He continued these studies first as a fellow and then as a faculty member at George Washington University from 1966 to 1973. Dr. Wright joined the faculty at Eastern Virginia Medical School (EVMS) in 1973, and became the department second Chairman in 1986. Since his arrival at EVMS, Dr. Wright's research has focused on the detection, identification, and characterization of biomarkers for early detection of urological cancers using a variety of molecular and proteomic technologies. In 1987, he cofounded the Virginia Prostate Center, a multidisciplinary program providing quality patient care, education, and research in urological diseases. The center's focus on proteomics resulted in the establishment of the Center for Biomedical Proteomics in 2001. Dr. Wright has published more than 150 original scientific papers and over 350 scientific presentations, and has 11 patents. He is a member of several scientific organizations, and has received numerous scientific awards; most recently (2000) the endowed EVMS Foundation Chair in Biomedical Science.

O. John Semmes received the PhD degree from George Washington University in 1989 in biochemistry. His PhD research addressed the structure/function of the high-affinity IL-2R. His postdoctoral studies at the National Institute of Health to investigate the molecular biology focused on the protein structure/function of human T-cell leukemia virus. Dr. Semmes joined the faculty at Johns Hopkins Medical School in 1995 as an Instructor. In 1997, he joined the University of Virginia as an Assistant Professor, and then as an Associate Professor at Eastern Virginia Medical School in 2000.
Dr. Semmes continues his research on HTLV-1utilizing both genomic and proteomic methodologies. He is currently the Director of the Center for Biomedical Proteomics, a program codeveloped with Dr. Wright, and is the Principal Investigator for the National Cancer Institutes' Early Detection Research Network Biomarker Discovery Laboratory at EVMS which focuses on biomarker discovery and early detection of prostate and breast cancers. He has published over 42 peer-reviewed scientific papers, given more than 150 scientific presentations, and had 4 pending patent applications. Dr. Semmes is a member of the HuPo Plasma group and a regular member of various NIH study sections and advisory groups concerned with cancer biology and proteomics.

* E-mail: wrightgl@evms.edu
† E-mail: semmesoj@evms.edu

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 $\times$ 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 post-translational 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

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

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

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

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:

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

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

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

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

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

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 $\mathbf{x}$ 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* (*k*NN)

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 ith gene in the mth 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 $\mathbf{x}$ is transformed and mapped to get a real-valued output $\mathbf{y}$:

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

where $\mathbf{W}$ is the weight matrix, called first layer, $\mathbf{h}$ is a nonlinear transformation, $\mathbf{y}$ 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

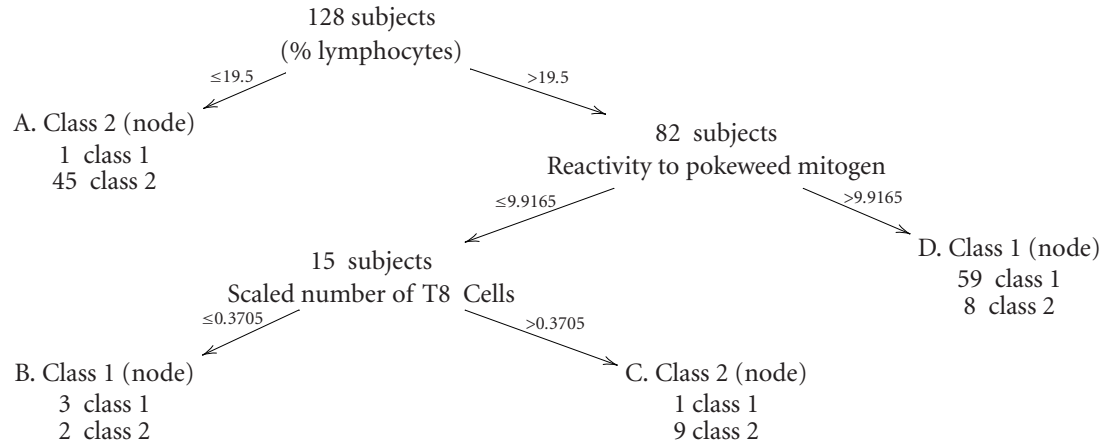


Figure 1. An example of neural network black box: a four-dimensional data input $\mathbf{x}$ is first transformed by $\mathbf{W}$, then by h in order to give a grouping variable $\mathbf{y}$ as an output.

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

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

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 log-normal 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 = $[\lambda DA_k D^t]$	Model 4 = $[\lambda_k DA_k D^t]$
Model 5 = $[\lambda D_k A D_k^t]$	Model 6 = $[\lambda_k D_k A D_k^t]$	Model 7 = $[\lambda D_k A_k D_k^t]$	Model 8 = $[\lambda_k D_k A_k D_k^t]$
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

Journal of Biomedicine and Biotechnology • 2003:4 (2003) 231–236 • PII. S1110724303209219 • http://jbb.hindawi.com

REVIEW ARTICLE

Bioinformatics Resources for In Silico Proteome Analysis

Manuela Pruess and Rolf Apweiler*

EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received 26 June 2002; accepted 10 December 2002

In the growing field of proteomics, tools for the in silico analysis of proteins and even of whole proteomes are of crucial importance to make best use of the accumulating amount of data. To utilise this data for healthcare and drug development, first the characteristics of proteomes of entire species—mainly the human—have to be understood, before secondly differentiation between individuals can be surveyed. Specialised databases about nucleic acid sequences, protein sequences, protein tertiary structure, genome analysis, and proteome analysis represent useful resources for analysis, characterisation, and classification of protein sequences. Different from most proteomics tools focusing on similarity searches, structure analysis and prediction, detection of specific regions, alignments, data mining, 2D PAGE analysis, or protein modelling, respectively, comprehensive databases like the proteome analysis database benefit from the information stored in different databases and make use of different protein analysis tools to provide computational analysis of whole proteomes.

INTRODUCTION

Continual advancement in proteome research has led to an influx of protein sequences from a wide range of species, representing a challenge in the field of Bioinformatics. Genome sequencing is also proceeding at an increasingly rapid rate, and this has led to an equally rapid increase in predicted protein sequences. All these sequences, both experimentally derived and predicted, need to be stored in comprehensive, nonredundant protein sequence databases. Moreover, they need to be assembled and analysed to represent a solid basis for further comparisons and investigations. Especially the human sequences, but also those of the mouse and other model organisms, are of interest for the efforts towards a better understanding of health and disease. An important instrument is the in silico proteome analysis.

The term "proteome" is used to describe the protein equivalent of the genome. Most of the predicted protein sequences lack a documented functional characterisation. The challenge is to provide statistical and comparative analysis and structural and other information for these sequences as an essential step towards the integrated analysis of organisms at the gene, transcript, protein, and functional levels.

Especially whole proteomes represent an important source for meaningful comparisons between species and furthermore between individuals of different health states. To fully exploit the potential of this vast quantity of data, tools for in silico proteome analysis are necessary. In the following, some important sources for proteome analysis like sequence databases and analysis tools will be described, which represent highly useful proteomics tools for the discovery of protein function and protein characterisation.

RESOURCES

Important tools for genome and proteome analysis are databases that store the huge amount of biological data, which is often no longer published in conventional publications. These databases, especially in combination with database search tools and tools for the computational analysis of the data, are necessary resources for biological and medical research.

Sequence databases

Sequence databases are of special importance for different fields of research because they are comprehensive sources of information on nucleotide sequences and proteins. There are basically three types of sequence-related databases, collecting nucleic acid sequences, protein sequences, and protein tertiary structures, respectively.

Nucleotide sequence databases

In nucleotide sequence databases, data on nucleic acid sequences as it results from the genome sequencing projects, and also from smaller sequencing efforts, is stored. The vast majority of the nucleotide sequence data produced is collected, organized, and distributed

by the International Nucleotide Sequence Database Collaboration [1], which is a joint effort of the nucleotide sequence databases EMBL-EBI (European Bioinformatics Institute, http://www.ebi.ac.uk), DDBJ (DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp), and GenBank (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov). The nucleotide sequence databases are data repositories, accepting nucleic acid sequence data from the community and making it freely available. The databases strive for completeness, with the aim of recording and making available every publicly known nucleic acid sequence. EMBL, GenBank, and DDBJ automatically update each other every 24 hours with new or updated sequences. Since their conception in the 1980s, the nucleic acid sequence databases have experienced constant exponential growth. There is a tremendous increase of sequence data due to technological advances. At the time of writing, the DDBJ/EMBL/GenBank Nucleotide Sequence Database has more than 10 billion nucleotides in more than 10 million individual entries. In effect, these archives currently experience a doubling of their size every year. Today, electronic bulk submissions from the major sequencing centers overshadow all other input and it is not uncommon to add to the archives more than 7000 new entries, on average, per day.

Protein sequence databases

In protein sequence databases, information on proteins is stored. Here it has to be distinguished between universal databases covering proteins from all species and specialised data collections storing information about specific families or groups of proteins, or about the proteins of a specific organism. Two categories of universal protein sequence databases can be discerned: simple archives of sequence data and annotated databases where additional information has been added to the sequence record. Especially the latter are of interest for the needs of proteome analysis.

PIR, the protein information resource [2] (http://www-nbrf.georgetown.edu/) has been the first protein sequence database which was established in 1984 by the National Biomedical Research Foundation (NBRF) as a successor of the original NBRF Protein Sequence Database. Since 1988 it has been maintained by PIR-International, a collaboration between the NBRF, the Munich Information Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID). The PIR release 71.04 (March 1, 2002) contains 283 153 entries. It presents sequences from a wide range of species, not especially focusing on human.

SWISS-PROT [3] is an annotated protein sequence database established in 1986 and maintained since 1988 collaboratively by the Swiss Institute of Bioinformatics (SIB) (http://www.expasy.org/) and the EMBL Outstation-The European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/swissprot/). It strives to provide a high level of annotation such as the description of the function of a protein, its domain structure, posttranslational modifications, variants, and so forth, and a minimal level of redundancy. More than 40 cross-references—about 4 000 000 individual links in total—to other biomolecular and medical databases, such as the EMBL/GenBank/DDBJ international nucleotide sequence database [1], the PDB tertiary structure database [4] or Medline, are providing a high level of integration. Human sequence entries are linked to MIM [5], the "Mendelian Inheritance in Man" database that represents an extensive catalogue of human genes and genetic disorders. SWISS-PROT contains data that originates from a wide variety of biological organisms. Release 40.22 (June 24, 2002) contains a total of 110 824 annotated sequence entries from 7459 different species; 8294 of them are human sequences. The annotation of the human sequences is part of the HPI project, the human proteomics initiative [6], which aims at the annotation of all known human proteins, their mammalian orthologues, polymorphisms at the protein sequence level, and posttranslational modifications, and at providing tight links to structural information and clustering and classification of all known vertebrate proteins. Seven hundred sixty-one human protein sequence entries in SWISS-PROT contain data relevant to genetic diseases. In these entries, the biochemical and medical basis of the diseases are outlined, as well as information on mutations linked with genetic diseases or polymorphisms, and specialised databases concerning specific genes or diseases are linked [7].

TrEMBL (translation of EMBL nucleotide sequence database) [3] is a computer-annotated supplement to SWISS-PROT, created in 1996 with the aim to make new sequences available as quickly as possible. It consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL nucleotide sequence database, except the CDSs already included in SWISS-PROT. TrEMBL release 21.0 (June 21, 2002) contains 671 580 entries, which should be eventually incorporated into SWISS-PROT, 32 531 of them human. Before the manual annotation step, automated annotation [8, 9] is applied to TrEMBL entries where sensible.

SP_TR_NRDB (or abbreviated SPTR or SWALL) is a database created to overcome the problem of the lack of comprehensiveness of single-sequence databases: it comprises both the weekly updated SWISS-PROT work release and the weekly updated TrEMBL work release. So SPTR provides a very comprehensive collection of human sequence entries, currently 45 629.

The *CluSTr* (clusters of SWISS-PROT and TrEMBL proteins) database [10] (http://www.ebi.ac.uk/clustr) is a specialised protein sequence database, which offers an automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences. Analysis has been carried out for different levels of protein similarity, yielding a hierarchical organisation of clusters.

Protein tertiary structure databases

The number of known protein structures is increasing very rapidly and these are available through *PDB*, the protein data bank [4] (http://www.rcsb.org/pdb/). There is also a database of structures of "small" molecules of interest to biologists concerned with protein-ligand interactions, available from the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk/).

In addition, there are also a number of derived databases, which enable comparative studies of 3D structures as well as to gain insight on the relationships between sequence, secondary structure elements, and 3D structure. *DSSP* (dictionary of secondary structure in proteins, http://www.sander.ebi.ac.uk/dssp/) [11] contains the derived information on the secondary structure and solvent accessibility for the protein structures stored in PDB. *HSSP* (homology-derived secondary structure of proteins, http://www.sander.ebi.ac.uk/hssp/) [12] is a database of alignments of the sequences of proteins with known structure with all their close homologues. *FSSP* (families of structurally similar proteins, http://www.ebi.ac.uk/dali/fssp/) [13] is a database of structural alignments of proteins. It is based on an all-against-all comparison of the structures stored in PDB. Each database entry contains structural alignments of significantly similar proteins but excludes proteins with high sequence similarity since these are usually structurally very similar.

The *SCOP* (structural classification of proteins) database [14] (http://scop.mrc-lmb.cam.ac.uk/scop/) has been created by manual inspection and abetted by a battery of automated methods. This resource aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds and detailed information about the close relatives of any particular protein.

Another database, which attempts to classify protein structures in the PDB, is the *CATH* database [15] (http://www.biochem.ucl.ac.uk/bsm/cath_new/), a hierarchical domain classification of protein structures in the PDB.

Proteome analysis databases and tools

Tools and databases for proteome analysis are based on reliable algorithms and information about protein sequences and structures derived from comprehensive protein databases. It can be difficult to distinguish between "database" and "tool" since databases providing precomputed data and search algorithms can offer a high functionality towards protein analysis.

Proteome analysis databases

The classic proteomics databases are those of 2D gel electrophoresis data such as the *SWISS-2DPAGE* database (two-dimensional polyacrylamide gel electrophoresis database) [16] (http:// www.expasy.ch/ch2d/). However, since the genome sequencing is proceeding at an increasingly rapid rate, this leads to an equally rapid increase in predicted protein sequences entering the protein sequence databases. Most of these predicted protein sequences are without a documented functional role. The challenge is to bridge the gap until functional data has been gathered through experimental research by providing statistical and comparative analysis and structural and other information for these sequences. This way of computational analysis can serve as an essential step towards the integrated analysis of organisms at the gene, transcript, protein, and functional levels.

Proteome analysis databases have been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms.

The *proteome analysis database* [17] (http://www.ebi.ac.uk/proteome) has the more general aim of integrating information from a variety of sources that will together facilitate the classification of the proteins in complete proteome sets. The proteome sets are built from the SWISS-PROT and TrEMBL protein sequence databases that provide reliable, well-annotated data as the basis for the analysis. Proteome analysis data is available for all the completely sequenced organisms present in SWISS-PROT and TrEMBL, spanning archaea, bacteria, and eukaryotes. In the proteome analysis effort, the InterPro [18] (http://www.ebi.ac.uk/interpro/) and CluSTr resources have been used. Links to structural information databases like the HSSP and PDB are provided for individual proteins from each of the proteomes. A functional classification using gene ontology (GO; [19]) is also available. The proteome analysis database provides a broad view of the proteome data classified according to signatures describing particular sequence motifs or sequence similarities and at the same time affords the option of examining various specific details like structural or functional classification. It currently (June 2002) contains statistical and analytical data for the proteins from 77 complete genomes.

The *international protein index* (IPI) (http://www.ebi.ac.uk/IPI/IPIhelp.html) provides a top-level guide to the main databases that describe the human and mouse proteome, namely SWISS-PROT, TrEMBL, RefSeq [20], and EnsEMBL [21]. IPI maintains a database of cross-references between the primary data sources with the aim of providing a minimally redundant yet maximally complete set of human proteins (one sequence per transcript).

Proteome analysis tools

Traditional proteomics tools like those accessible from the ExPASy server (http://www.expasy.org) represent a variety of possibilities to analyse proteins. They help to identify and characterise proteins, to convert DNA sequences into amino acid sequences, and to perform similarity searches, pattern and profile searches, post-translational modification prediction, primary structure

Table 1. InterPro comparative analysis of *Homo sapiens* and *Mus musculus* proteomes—the first 17 of the top 30 hits are shown.

	H sapiens		*M musculus*		
InterPro	Proteins matched (Proteome coverage)	Rank	Proteins matched (Proteome coverage)	Rank	Description
IPR000822	1165 (3.4%)	1	341 (1.4%)	5	Zn-finger, C2H2 type
IPR003006	928 (2.7%)	2	498 (2.1%)	3	Immunoglobulin/major histocompatibility complex
IPR000719	738 (2.2%)	3	387 (1.6%)	4	Eukaryotic protein kinase
IPR000694	713 (2.1%)	4	0		Poline-rich region
IPR000276	681 (2.0%)	5	401 (1.7%)	29	Rhodopsin-like GPCR superfamily
IPR002290	515 (1.5%)	6	275 (1.1%)	7	Serine/threonine protein kinase
IPR000561	417 (1.2%)	7	212 (0.9%)	13	EGF-like domain
IPR001909	405 (1.2%)	8	85 (0.4%)	15	KRAB box
IPR001680	386 (1.1%)	9	168 (0.7%)	17	G-protein beta WD-40 repeat
IPR001245	375 (1.1%)	10	180 (0.7%)	10	Tyrosine protein kinase
IPR001841	358 (1.1%)	11	180 (0.7%)	34	Zn-finger, RING
IPR003599	347 (1.0%)	12	174 (0.7%)	8	Immunoglobulin subtype
IPR000504	346 (1.0%)	13	156 (0.6%)	21	RNA-binding region RNP-1 (RNP recognition motif)
IPR003600	345 (1.0%)	14	170 (0.7%)	6	Immunoglobulin-like
IPR001849	326 (1.0%)	15	128 (0.5%)	33	Pleckstrin-like
IPR002965	299 (0.9%)	16	102 (0.4%)	11	Proline-rich extensin
IPR001452	296 (0.9%)	17	138 (0.6%)	32	SH3 domain

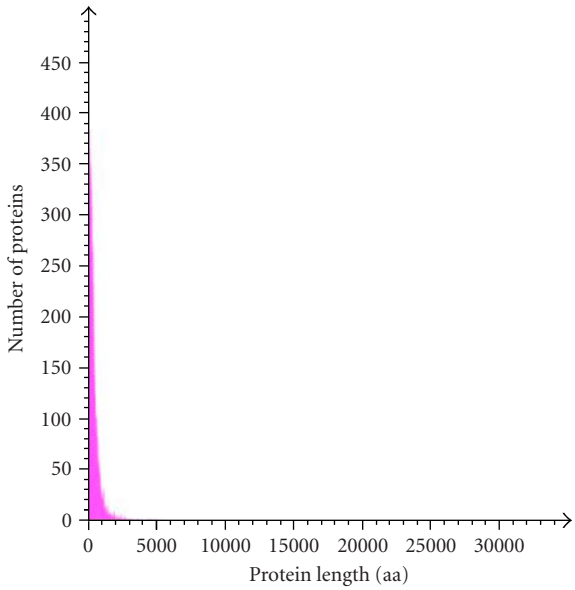
analysis, secondary and tertiary structure prediction, detection of transmembrane regions, alignments, and biological text analysis. Moreover, there is a software available for 2D PAGE analysis, automated knowledge-based protein modelling, and structure display and analysis.

The analysis of whole proteomes represents an even bigger challenge. Large and comprehensive databases and knowledge bases are developed and used which provides large sets of precomputed data. To gather this comprehensive data, a vast amount of underlying information is necessary. The *proteome analysis database*, mentioned above, uses annotated information about proteins from the SWISS-PROT/TrEMBL database and automated protein classifications from InterPro, CluSTr, HSSP, TMHMM [22], and SignalP [23]. The precomputation permits comparisons of whole proteomes of completely sequenced organisms with those of others (Table 1). Users of the database can perform their own interactive proteome comparisons between any combinations of organisms in the database. Moreover, structural features of individual proteomes like the protein length distribution (Figure 1), amino acid composition (Figure 2), affiliation of the different proteins to protein families, and the number of sequences in total and those displayed by other databases can be requested. Users are also able to run a Fasta similarity search (Fasta3) on their own sequence against a complete proteome in the database with the help of a specific search form. It is possible to download a proteome set or a list of InterPro matches for a given organism, to see the current status of all complete proteomes in SWISS-PROT and TrEMBL, and to download GO annotation for the human proteome.

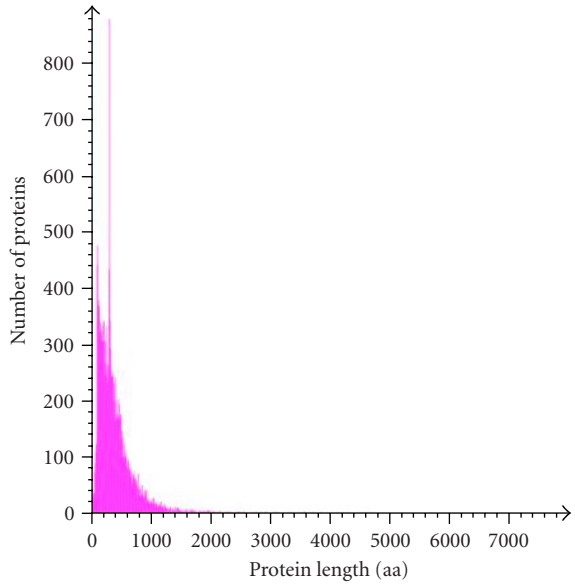
Other tools important especially for laboratory scientists are image analysis tools, laboratory information management systems, and software for the characterisation from mass spectrometric data.

DISCUSSION

In the last years there has been a tremendous increase in the amount of data available concerning the human genome and more particularly the molecular basis of genetic diseases. Every week, new discoveries are made that link one or more genetic diseases to defects in specific genes. To take into account these developments, the SWISS-PROT protein sequence database for example is gradually enhanced by the addition of a number of features that are specifically intended for researchers working on the basis of human genetic diseases as well as the extent of polymorphisms. The latter are very important too, since they may represent the basis for differences between individuals, which are particularly interesting

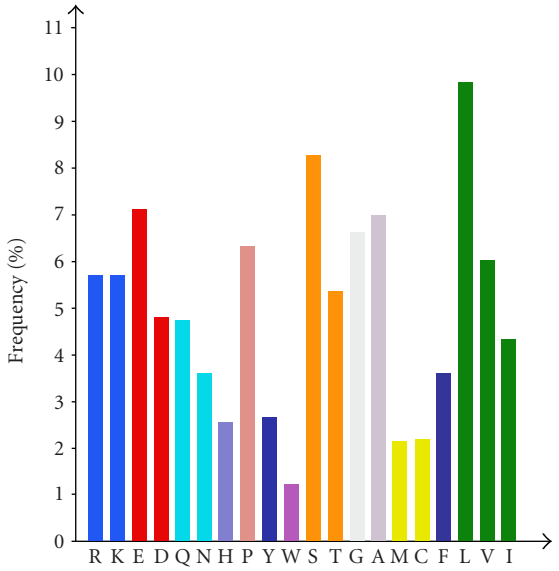


(a) *Homo sapiens*. Analysis of full-length proteins (fragments excluded). Average proteins length: 469 ± 567 amino acid residues.

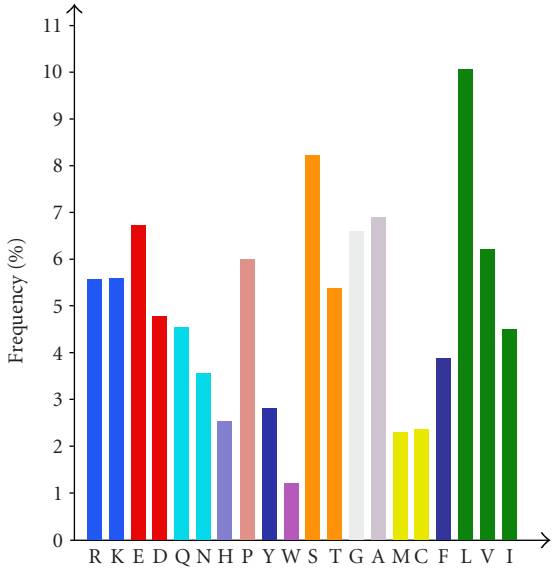


(b) *Mus musculus* (mouse). Analysis of full-length proteins (fragments excluded). Average proteins length: 416 ± 384 amino acid residues. Size range: 10–7389 amino acid residues.

FIGURE 1. Protein length distribution of *Homo sapiens* and *Mus musculus*.



(a) *Homo sapiens*



(b) *Mus musculus* (mouse)

FIGURE 2. Amino acid composition of *Homo sapiens* and *Mus musculus*. (The total number of each amino acid in each proteome is given additionally as well as the frequency in (%).)

for some aspects of medicine and drug research. Such comprehensive sequence databases are mandatory for the use of proteome analysis tools like the proteome analysis database, which combines the different protein sequences of a given organism to a complete proteome. This proteome can be regarded as a whole new unit, analysable according to different points of view (like distribution of domains and protein families, and secondary and tertiary structures of proteins), and can be made comparable to other proteomes. In general, for using the proteomics data for healthcare and drug development, first the characteristics of proteomes of entire species—mainly the human—have to be understood before secondly differentiation between individuals can be surveyed.

But although the number of proteome analysis tools and databases is increasing and most of them are providing a very good quality of computational efforts and/or annotation of information, the user should not forget that automated analysis always can hold some mistakes. Data material in databases is reliable, but only to a certain point. Automatic tools which use data derived from databases can thus be error-prone, rules built on their basis can be wrong, and sequence similarities can occur due to chance and not due to relationship. Users of bioinformatics tools should in no way feel discouraged in their using, they only should keep in mind the potential pitfalls of automated systems and even of humans—and be encouraged to check all data as far as possible and not blindly rely on them.

REFERENCES

[1] Stoesser G, Baker W, van den Broek A, et al. The EMBL nucleotide sequence database. *Nucleic Acids Res*. 2001;29(1):17–21.

[2] Wu CH, Huang H, Arminski L, et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res*. 2002;30(1):35–37.

[3] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 2000;28(1):45–48.

[4] Bhat TN, Bourne P, Feng Z, et al. The PDB data uniformity project. *Nucleic Acids Res*. 2001;29(1):214–218.

[5] Pearson PL, Francomano C, Foster P, Bocchini C, Li P, McKusick VA. The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res*. 1994;22(17):3470–3473.

[6] O'Donovan C, Apweiler R, Bairoch A. The human proteomics initiative (HPI). *Trends Biotechnol*. 2001;19(5):178–181.

[7] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med*. 1997;75(5):312–316.

[8] Fleischmann W, Moller S, Gateau A, Apweiler R. A novel method for automatic functional annotation of proteins. *Bioinformatics*. 1999;15(3):228–233.

[9] Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*. 2001;17(10):920–926.

[10] Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res*. 2001;29(1):33–36.

[11] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577–2637.

[12] Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res*. 1998;26(1):313–315.

[13] Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*. 1996;24(1):206–209.

[14] Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*. 2002;30(1):264–267.

[15] Pearl FMG, Martin N, Bray JE, et al. A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res*. 2001;29(1):223–227.

[16] Hoogland C, Sanchez JC, Tonella L, et al. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res*. 2000;28(1):286–288.

[17] Apweiler R, Biswas M, Fleischmann W, et al. Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res*. 2001;29(1):44–48.

[18] Apweiler R, Attwood TK, Bairoch A, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*. 2001;29(1):37–40.

[19] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–29.

[20] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*. 2001;29(1):137–140.

[21] Hubbard T, Barker D, Birney E, et al. The Ensemble genome database project. *Nucleic Acids Res*. 2002;30(1):38–41.

[22] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In: Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C, eds. *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, Calif: AAAI Press;1998:175–182.

[23] Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng*. 1999;12(1):3–9.

* Corresponding author.
E-mail: apweiler@ebi.ac.uk
Fax: +44 1223 49 44 68;
Tel: +44 1223 49 44 35

Journal of Biomedicine and Biotechnology • 2003:4 (2003) 237–241 • PII. S1110724303210020 • http://jbb.hindawi.com

REVIEW ARTICLE

SELDI ProteinChip® Array Technology: Protein-Based Predictive Medicine and Drug Discovery Applications

Guru Reddy and Enrique A. Dalmasso*

Ciphergen Biosystems, Inc, 6611 Dumbarton Circle, Fremont, CA 94555, USA

Received 24 October 2002; accepted 5 November 2002

Predictive medicine, utilizing the ProteinChip® Array technology, will develop through the implementation of novel biomarkers and multimarker patterns for detecting disease, determining patient prognosis, monitoring drug effects such as efficacy or toxicity, and for defining treatment options. These biomarkers may also serve as novel protein drug candidates or protein drug targets. In addition, the technology can be used for discovering small molecule drugs or for defining their mode of action utilizing protein-based assays. In this review, we describe the following applications of the ProteinChip Array technology: (1) discovery and identification of novel inhibitors of HIV-1 replication, (2) serum and tissue proteome analysis for the discovery and development of novel multimarker clinical assays for prostate, breast, ovarian, and other cancers, and (3) biomarker and drug discovery applications for neurological disorders.

INTRODUCTION

The ProteinChip Array technology is used for the discovery, validation, identification, and characterization of disease-associated proteins from biological samples. The versatile nature of this technology is enabling for a wide variety of applications in both research and clinical proteomics and will be reviewed in three application areas. A recent publication presents the use of the flexibility and power of the ProteinChip Array platform to elucidate the nature of novel protein inhibitors of HIV-1 replication, the molecules previously known as the $CD8^+$ antiviral factors. Numerous cancer-related publications have demonstrated the discovery and development of biomarkers and multimarker patterns for protein-based predictive medicine. Finally, we discuss a variety of drug discovery applications using Alzheimer's disease as the model system.

SELDI PROTEINCHIP® ARRAY TECHNOLOGY

The ProteinChip Array technology is based on surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) [1]. The key components of this technology are the ProteinChip Arrays, the ProteinChip Reader, and the associated software. ProteinChip Arrays use various chromatography surfaces or biological surfaces to capture proteins from complex biological mixtures according to their physicochemical properties. Chromatographic surfaces are composed of hydrophobic, hydrophilic, ion exchange, immobilized metal, or other chemistries. These surfaces are often used for profiling proteins from biological mixtures, for biomarker discovery, and for assay implementation. This is considered a *de novo* protein discovery approach in that no prior knowledge of the specific proteins is required as would be the case using antibody-based arrays. The activated surfaces are used to covalently immobilize specific bait molecules such as antibodies, receptors, or oligonucleotides often used for biomolecular interaction studies such as protein-protein and protein-DNA interactions.

Biological samples such as cell lysates, tissue extracts, or biological fluids are added to a spot on a ProteinChip Array and proteins are allowed to bind to the surface based on the general chromatographic or specifically designed biological affinity properties. The unbound proteins and mass spectrometric interfering compounds are washed away and the proteins that are retained on the array surface are analyzed and detected by SELDI-TOF-MS using a ProteinChip Reader (Figure 1). The MS profiles from the various sets of samples are then compared in a technique described as differential protein expression mapping, whereby relative expression levels of proteins at specific molecular weights are compared by a variety of statistical techniques and bioinformatic software systems [2].

DISCOVERY AND IDENTIFICATION OF HIV REPLICATION INHIBITORS

HIV has so far infected 40 million people and 20 millions have died of the AIDS disease. It is expected that

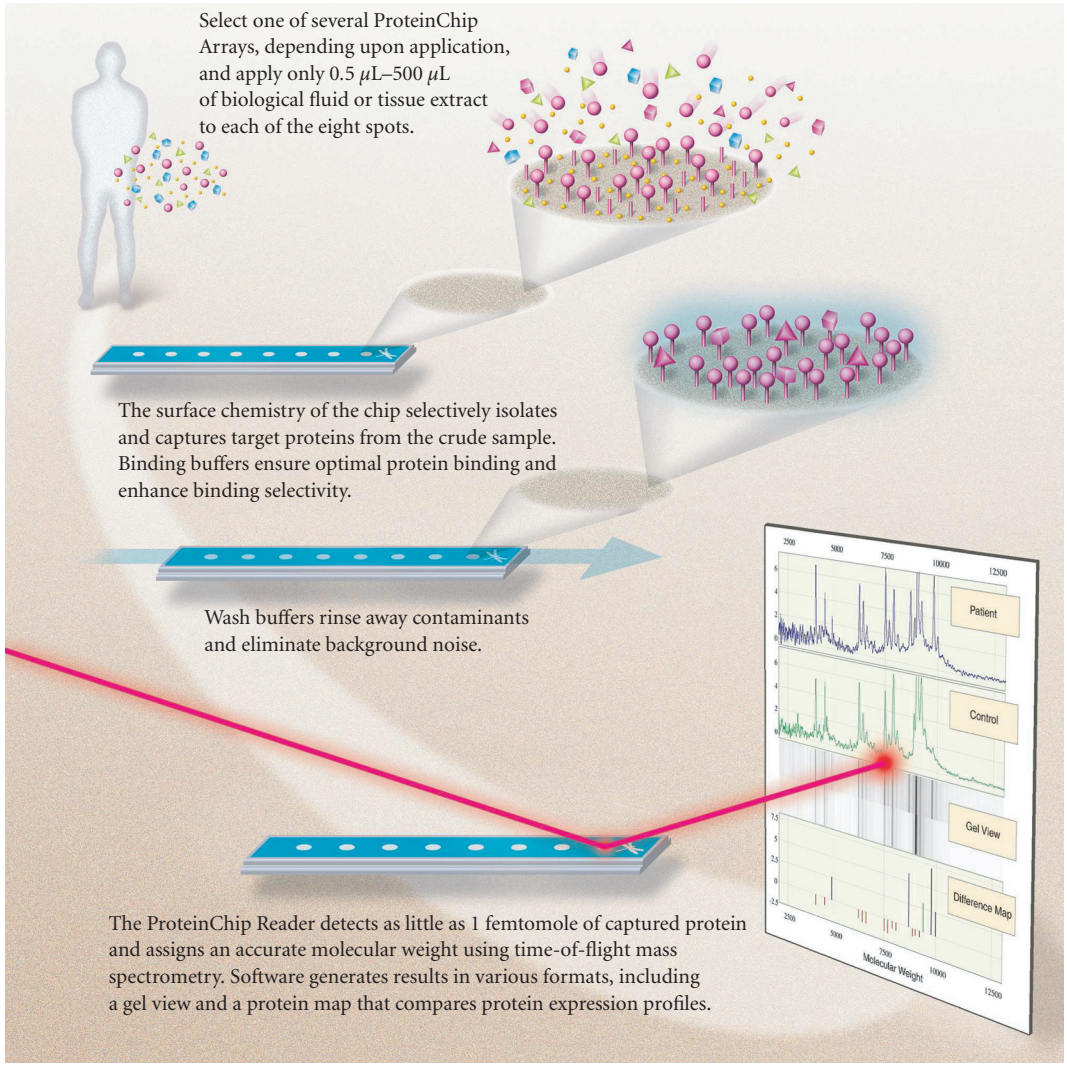


FIGURE 1. Protein profiling protocol. Procedure for preparing the ProteinChip Arrays with biological samples and for analyzing the retained proteins by SELDI-TOF-MS using a ProteinChip Reader.

around 3 millions more deaths will occur over the next 12 months, equating to over 9,000 deaths per day. Every day, 15,000 more people are infected. Certain individuals infected with HIV-1 virus remain in clinically stable condition for many years after the infection and are classified as long-term nonprogressors (LTNP). It has been known for some time that $CD8^+$ T lymphocytes from these individuals secrete a soluble factor, $CD8^+$ antiviral factor (CAF), that suppresses HIV-1 replication, irrespective of viral phenotype [3]. CAF is found in greater abundance in supernatants from LTNP $CD8^+$ cells and not usually detected from patients that progress to develop AIDS.

Despite enormous efforts by numerous laboratories over the past 16 years, the identity of CAF has remained elusive. Traditional approaches, including protein expression based on mRNA levels and several proteomic technologies, had failed to define the molecules responsible for the full extent of this activity. Several studies have shown that CAF lacks full identity to known chemokines. Interleukin 16 (IL-16) was suggested as the identity of CAF by Kruth and colleagues [4]; however, it is not present in all CAF-containing fluids and it only reduces HIV replication in an acute infection assay at high concentrations. Hence, it does not account for the CAF activity. It was postulated by Gallo and colleagues that beta-chemokines such as RANTES, MIP-1-alpha, and MIP-1-beta were responsible for the antiviral activity of $CD8^+$ cells [5]. But later it was shown that beta-chemokines can competitively block the R5 viruses that use CCR5 as a coreceptor but not the X4 viruses that use CXCR4 as a coreceptor. Since CAF can inhibit both types of viruses, beta-chemokines do not fully account for CAF activity.

Drs David Ho and Linqi Zhang and their colleagues from the Aaron Diamond AIDS Research Center were supported by the Ciphergen Biomarker Discovery Center® team in their discovery of the identity of CAF.

Using Ciphergen's ProteinChip System in their laboratory, Ho and colleagues discovered a cluster of low molecular weight proteins (3.3 to 3.5 kd) that were present in stimulated $CD8^+$ cells from normal individuals and LTNP, but not from patients progressing to AIDS. These unique proteins were enriched by ion exchange followed by reverse-phase chromatography. "This enrichment and purification process was monitored by SELDI-TOF-MS." Following enrichment, these proteins were then definitively identified as human alpha-defensin-1, -2, and -3 using Ciphergen's PCI-1000 ProteinChip Interface on a tandem MS system [6].

These findings were confirmed by in vitro inhibition of viral replication for many HIV-1 isolates. Antibodies specific for human alpha-defensins completely eliminated the CAF activity, while the specific antibody-based depletion of the molecules was demonstrated by an on-chip SELDI-TOF-MS assay. Furthermore, the authors demonstrated that commercially synthesized alpha-defensins also inhibit the in vitro replication of HIV-1. Taken together, all of these results indicated that alpha-defensin-1, -2, and -3 collectively account for most of the anti-HIV activity of CAF. Detailed mechanism and in vivo studies will define the modes of action and biological utility of the alpha-defensins. This discovery will have a profound impact on the development of novel AIDS therapeutic and diagnostic approaches. This is a powerful example demonstrating that in addition to the *de novo* discovery of biomarker and multimarker patterns, SELDI ProteinChip Array technology can identify novel protein therapeutic drug candidates and assign novel biological function to known proteins.

MULTIMARKER CLINICAL ASSAYS FOR CANCER

Cancer is the second leading cause of death in the Untied States, exceeded only by heart disease. In the USA, one out of every four deaths is from cancer. This year 555,000 Americans are expected to die of cancer, more than 1500 people a day. Discovery and development of better diagnostics and therapies are urgently needed to fight this deadly disease. Most current diagnostic tests detect cancer in advanced stages when treatment is often difficult and prognosis is poor. Numerous studies have shown that early detection of cancer increases treatment options and improves survival rates. Novel biomarkers can be discovered by comparing the differences in protein expression profiles between serum or tissue extract samples from cancer patients and normal individuals. SELDI ProteinChip Array technology has been used extensively to profile proteins and discover biomarkers in different types of cancers [7, 8, 9].

The laboratories of Drs George Wright, Jr, Daniel Chan, Lance Liotta, Emanuel Petricoin, and many others are currently using the SELDI ProteinChip Array technology for serum proteome analysis. These laboratories are focused on the discovery of markers and biomarker patterns for the early detection of prostate, breast, ovarian, and other cancers. The main objectives of these studies are to find signature proteomic patterns, or multimarker clinical assays, in serum that differentiate normal individuals from cancer patients and have clinical performance better than current single markers, thereby enabling more accurate diagnoses by accounting for the heterogeneity of cancer. In these studies, protein profiling data is generated by SELDI ProteinChip Array technology followed by analysis utilizing numerous types of multivariate software algorithms.

Research from these laboratories utilizing hundreds of serum samples per study has led to the discovery of multiple biomarkers or biomarker patterns that, when used in combination, have higher clinical sensitivity and specificity relative to the best available single-marker assays. For example, Wright and colleagues [7] discovered that a SELDI multimarker profile combining nine different proteins generates an assay with better sensitivity (83%) and specificity (97%) for diagnosis of prostate cancer than the prostate specific antigen (PSA) test. Chan and colleagues [8] demonstrated that when three newly discovered biomarkers for breast cancer are used in combination, the SELDI assay has a significantly higher sensitivity (93%) and specificity (91%) relative to CA15.3, the best available protein marker. For ovarian cancer, Liotta and Petricoin et al [9] demonstrated that a SELDI multimarker profile has sensitivity (100%) and specificity (95%), compared to the poor performance of the CA125 test.

While studying serum protein profiles is helpful in the early diagnosis of cancer, the final confirmation and staging of the disease comes from examining the tumor tissue itself. Since tumor tissue is heterogeneous in nature, new molecular techniques are needed to study the biomarkers that are present. Laser capture microdissection (LCM) developed by Liotta and colleagues at the National Institutes of Health has enabled researchers to selectively procure pure population of cells from a stained tissue section under direct microscopic visualization. The comprehensive analysis of several cancer samples using SELDI-TOF-MS to generate tissue-specific profiles of LCM-procured samples has been reported [10, 11]. Wright and colleagues [11] used ProteinChip Array technology to discover several protein biomarkers, in cells procured by LCM, that were specifically distinguished prostate cancer cells from surrounding normal prostate cells from the same patient. Pawaletz et al [10] analyzed protein profiles from patient-matched normal, premalignant, malignant, and metastatic microdissected cell populations from human esophageal, prostate, breast, ovary, colon, and hepatic tissue sections by SELDI-TOF-MS. They obtained reproducible and discriminatory protein biomarker profiles that differentiated normal cells from tumor cells and discriminated different tumor types. Coupling LCM for specific cell procurement with SELDI ProteinChip Array technology has a tremendous potential for the discovery

of specific biomarkers that are associated with each stage of tumor development.

Both serum and tissue proteome analysis by SELDI ProteinChip Array technology will have great utility in protein-based predictive medicine as demonstrated in these studies. Detailed molecular analysis of tissue has great potential to assist with improved definition of tumor aggressiveness and patient prognosis, and to assist with selection of appropriate treatment option. Also, SELDI ProteinChip Array multimarker serum protein patterns appear to perform significantly better in diagnosing different types of cancers than currently used single-marker assays. If these multimarker SELDI-TOF-MS protein profiles are validated in larger and more clinically diverse study sets, this approach can have immediate and substantial benefit for the early detection of many kinds of cancers. As these studies expand, applications for protein-based predictive medicine will further develop by establishing multimarker serum assays that address the more defined clinical questions described above but with a more easily acquired biological sample, possibly when the solid tumor itself cannot be located.

BIOMARKER AND DRUG DISCOVERY APPLICATIONS IN NEUROLOGICAL DISORDERS

The SELDI ProteinChip technology has been used by several researchers to discover novel biomarkers and multimarker panels that are relevant to the diagnosis and patient stratification for Alzheimer's disease, Parkinson's disease, multiple sclerosis, schizophrenia, HIV-induced dementia, and so forth. In this section, we focus on drug discovery applications of SELDI ProteinChip technology in Alzheimer's disease (AD). As the most prevalent form of neurodegenerative disorders, AD affects about 4 million people in the USA, generally people over the age of 65. Clinical features of AD include beta-amyloid deposits in brain, neuritic plaques, and degeneration of synapses [12]. In the familial forms of AD, mutation in the amyloid precursor protein (APP) or in a presenilin gene (PS1 or PS2) leads to increased amounts of the 40- and 42-amino acid beta-amyloid peptides that are primary components of plaques. The peptides originate from the proteolysis of APP by two proteases known as beta-secretase and gamma-secretase [13]. The advantage of using ProteinChip technology for the analysis is that they can be detected directly from a wide range of samples including cell culture supernatant, CSF, and brain/nerve lysates. Moreover, very low amounts of sample are required.

All of the beta-amyloid peptides share a common N-terminal sequence, so an antibody that is specifically raised against this N-terminal sequence can be immobilized on the ProteinChip Array and used to capture beta-amyloid peptides from complex samples. The SELDI-TOF-MS analysis of such a capture can be used to monitor the relative amounts of peptides of various lengths, including many from beta-amyloid 1–15 to 1–42, some of which correlate more strongly with the development of AD than others. This assay has been used successfully to profile peptide variants secreted into the media of cultured human neuronal cells that express the APP [14]. In addition, this assay was used to establish that BACE-1 is the beta-secretase that is involved in the generation of beta-amyloid peptides by neurons [15] and to discover candidate biomarkers for AD from human CSF (unpublished results, 2002). Secretases are thought to be potential drug targets and this assay has been widely used to discover novel secretase inhibitors [16], to study the function of secretase inhibitors [17], and to monitor the changes of beta-amyloid in serum and brain mediated by beta-amyloid vaccination [18]. The drug discovery capabilities of SELDI ProteinChip technology have been best demonstrated in the various uses of this on-chip assay for monitoring relative changes in various lengths of the beta-amyloid peptides.

DISCUSSION

In this review, we have described a number of recent publications that demonstrate the power of the SELDI ProteinChip Array technology when utilized directly by the translational medicine clinical researcher. The technology enables rapid testing of a clinical hypothesis, which greatly enhances and accelerates discovery potential and will enable protein-based predictive medicine. In the first example, demonstration of alpha-defensin-1, -2, and -3 as novel inhibitors of HIV-1 replication, the technology was instrumental in discovering and identifying a set of small proteins that had eluded researchers for 16 years. In the second set of examples, we summarized results from several publications presenting serum and tissue proteome analyses. These studies describe the discovery and development of multimarker clinical assays for prostate, breast, ovarian, and other cancers. These assays hold great promise for the early detection of cancer and for development of protein-based predictive medicine. Finally, in the third set of examples, we described publications that show the utility of the technology as demonstrated by drug discovery applications for AD. In addition to novel applications for protein-based predictive medicine, the technology is powerful in its ability to discover and characterize novel protein drug candidates, protein-protein interactions, and signal transduction pathways in the tumor cells which would in turn be used to customize therapy specific for each individual.

REFERENCES

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

[2] Fung ET, Enderwick C. ProteinChip® clinical proteomics: computational challenges and solutions. *Biotechniques*. 2002;32(suppl):34–41.

[3] Walker CM, Moody DJ, Stites DP, Levy JA. $CD8^+$ lymphocytes can control HIV infection in vitro by suppressing virus replication. *Science*. 1986;234(4783):1563–1566.
[4] Baier M, Werner A, Bannert N, Metzner K, Kurth R. HIV suppression by interleukin-16. *Nature*. 1995; 378(6557):563.
[5] Cocchi F, DeVico AL, Garzino-Demo A, Arya SK, Gallo RC, Lusso P. Identification of RANTES, MIP-1 alpha, and MIP-1 beta as the major HIV-suppressive factors produced by $CD8^+$ T cells. *Science*. 1995;270(5243):1811–1815.
[6] Zhang L, Yu W, He T, et al. Contribution of human α-defensin 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor. *Science Express*. 2002;298:995–1000.
[7] 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.
[8] 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.
[9] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–577.
[10] Paweletz CP, Gillespie JW, Ornstein DK, et al. Rapid protein display profiling of cancer progrssion directly from human tissue using a protein biochip. *Drug Development Research*. 2000;49:34–42.
[11] 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.
[12] Wisniewski HM, Terry RD. Reexamination of the pathogenesis of the senile plaque. In: Zimmerman HM, ed. *Progress in Neuropathology*. New York, NY: Grune and Stratton; 1973:1–26.
[13] Selkoe DJ. Alzheimer's disease: genotypes, phenotypes, and treatments. *Science*. 1997;275(5300):630–631.
[14] Davies H, Lomas L, Austen B. Profiling of amyloid β peptide variants using SELDI ProteinChip® arrays. *Biotechniques*. 1999;27(6):1258–1261.
[15] Cai H, Wang Y, McCarthy D, et al. BACE1 is the major β-secretase for generation of Aβ peptides by neurons. *Nat Neurosci*. 2001;4(3):233–234.
[16] Shearman MS, Beher D, Clarke EE, et al. L-685, 458, an aspartyl protease transition state mimic, is a potent inhibitor of amyloid β-protein precursor gamma-secretase activity. *Biochemistry*. 2000;39(30):8698–8704.
[17] Vandermeeren M, Geraerts M, Pype S, Dillen L, van Hove C, Mercken M. The functional γ-secretase inhibitor prevents production of amyloid β 1-34 in human and murine cell lines. *Neurosci Lett*. 2001;315(3):145–148.
[18] Vehmas AK, Borchelt DR, Price DL, et al. β-Amyloid peptide vaccination results in marked changes in serum and brain A β levels in APPswe/PS1 Δ E9 mice, as detected by SELDI-TOF-based ProteinChip® technology. *DNA Cell Biol*. 2001;20(11):713–721.

* Corresponding author.
E-mail: edalmasso@ciphergen.com
Tel: +1 510 505 2245

Journal of Biomedicine and Biotechnology • 2003:4 (2003) 242–248 • PII. S111072430320927X • http://jbb.hindawi.com

RESEARCH ARTICLE

An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers

Yutaka Yasui,[1]* Dale McLerran,[1] Bao-Ling Adam,[2] Marcy Winget,[1] Mark Thornquist,[1] and Ziding Feng[1]

[1]*Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MP 702, Seattle, WA 98109-1024, USA*
[2]*Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, 700 Olney Road, Norfolk, VA 23507, USA*

Received 25 July 2002; revised 20 December 2002; accepted 20 December 2002

Discovery of "signature" protein profiles that distinguish disease states (eg, malignant, benign, and normal) is a key step towards translating recent advancements in proteomic technologies into clinical utilities. Protein data generated from mass spectrometers are, however, large in size and have complex features due to complexities in both biological specimens and interfering biochemical/physical processes of the measurement procedure. Making sense out of such high-dimensional complex data is challenging and necessitates the use of a systematic data analytic strategy. We propose here a data processing strategy for two major issues in the analysis of such mass-spectrometry-generated proteomic data: (1) separation of protein "signals" from background "noise" in protein intensity measurements and (2) calibration of protein mass/charge measurements across samples. We illustrate the two issues and the utility of the proposed strategy using data from a prostate cancer biomarker discovery project as an example.

INTRODUCTION

With recent advances in mass spectrometry technologies, it is now possible to study protein profiles over a wide range of molecular weights in small biological specimens [1]. A key research step in translating these technological advancements into clinical utilities is the identification of "signature" protein profiles that distinguish disease states (eg, malignant, benign, and normal) or experimental conditions (eg, treated versus untreated by a drug of interest). For example, a discovery of disease-specific protein profiles could facilitate early detection of the disease and, consequently, contribute importantly towards improving patients' prognosis and survival.

Protein data generated from mass spectrometers have complex features due to complexities in both biological specimens and interfering biochemical/physical processes of the measurement procedure. They are also large in size, generally in the order of tens of thousands of measurement points per sample. Making sense out of such high-dimensional complex data is challenging and necessitates the use of a systematic data analytic strategy. Specifically, there are three major issues in the analysis of mass-spectrometry-generated protein data that need to be resolved effectively by the systematic strategy: (1) separation of protein "signals" from background "noise" in protein intensity measurements; (2) calibration of protein mass/charge measurements across samples; and (3) construction of "signature profiles," as combinations of multiple mass/charge points, that distinguish disease states or experimental conditions.

This paper is concerned with the first two of the three issues in the analysis of mass-spectrometry-generated protein data. We propose a systematic data processing method for separating signals (protein intensity peaks) from noise, and for calibrating mass/charge values of proteins that may fluctuate slightly at random across samples; for approaches to the third data analytic problem, several approaches have been proposed [2, 3, 4, 5, 6].

MATERIALS AND METHODS

In this section, we first describe the prostate cancer biomarker discovery project [4, 5] to illustrate the type of research settings of interest. The project's data are used to explain the signal identification and calibration issues. We then present our proposed data processing method that addresses these issues.

The prostate cancer biomarker discovery project

The Department of Microbiology and Molecular Cell Biology and Virginia Prostate Center of the Eastern Virginia Medical School (EVMS) have been conducting a biomarker discovery project on prostate cancer with a goal of identifying serum protein biomarkers of the

disease. This project is part of a large research consortium, the Early Detection Research Network [7, 8], funded by the National Cancer Institute. The basis for the protein-based early detection of cancer is the concept that a transformed cancerous cell and its clonal expansion would result in up- (or down-) regulation of certain proteins; our aim is to identify such early molecular signals of prostate cancer by measuring protein profiles in serum.

In this project, serum samples of 386 subjects were retrieved from the serum repository of the EVMS Virginia Prostate Center, approximately equally from four disease groups: late-stage prostate cancer ($N = 98$); early-stage prostate cancer ($N = 99$); benign prostatic hyperplasia (BPH) ($N = 93$); and normal controls ($N = 96$). The four disease groups were defined as follows: prostate cancer cases had a positive biopsy that was staged A or B (early-stage) or C or D (late-stage) and had a prostate specific antigen (PSA) concentrations greater than 4 ng/ml; BPH patients had PSA values between 4 ng/ml and 10 ng/ml, low PSA velocities, and at least two negative biopsies; and normal controls were aged 50 or older (ie, the same age range as cancer and BPH patients), had a PSA level less than 4 ng/ml, and had a normal digital rectal exam.

Each of the retrieved serum samples was assayed at the EVMS for protein expression by the surface-enhanced laser desorption/ionization (SELDI) ProteinChip Array technology [1, 2, 3, 4, 8, 9, 10, 11, 12, 13] of Ciphergen Biosystems, Inc, 6611 Dumbarton Circle, Fremont, CA 94555. The SELDI technology is a time-of-flight mass spectrometry with a special ProteinChip® Array whose surface captures proteins using chemically or biologically defined protein-docking sites. Proteins are captured on the chip surface, purified by washing the surface, and crystallized with small molecules called "matrix" or "energy-absorbing molecules" whose function is to absorb laser energy and transfer it to proteins. Energized protein molecules fly away from the surface into a time-of-flight tube where the time for the molecules to fly through the tube is a function of the molecular weight and charge of the protein. A detector at the end of the tube measures the "intensity" of proteins at each discrete time of flight and outputs about 48 000 data points of time of flight, intensity pairs. Each discrete time of flight corresponds uniquely to a ratio of the molecular weight of a protein to the number of charges introduced by the ionization. SELDI output, therefore, produces about 48 000 data points of mass/charge, intensity pairs. Our analyses used 16 898 data points per sample covering the mass/charge range of 2 000–40 000. Figure 1 shows an example of SELDI output from the first subject in the normal control group of the prostate cancer biomarker discovery project.

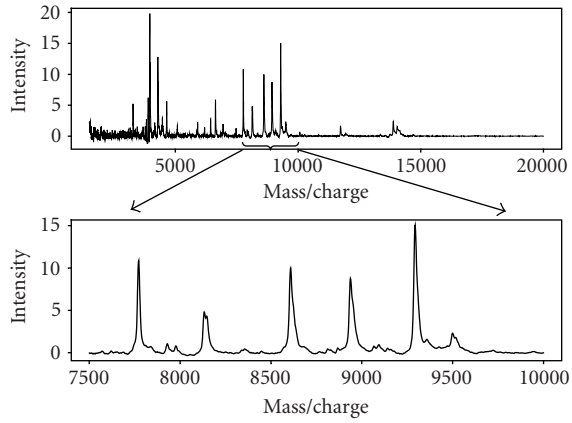


Figure 1. An example of SELDI output: the top panel shows the protein intensity measures (y-axis) in the range of mass/charge values below 20, 000 (x-axis) and the bottom panel zooms into a subregion covering 7 500–10 000 mass/charge values.

The two data analytic issues

The first data analytic issue is the mathematical definition of "peaks" in protein intensity, that is, the identification of "signals" separated from "noise" in protein intensity measurements. Defining protein intensity peaks mathematically provides two advantages. First, it reduces the dimensionality of data from tens of thousands of data points to a more manageable size (eg, less than a thousand data points). Second, perhaps more importantly, it clarifies the interpretation of "signature" protein profiles, the end products of the analysis. Specifically, protein intensity peaks and their heights at certain mass/charge values indicate the presence and the approximate amount of corresponding proteins or peptides in the specimen being analyzed. Without this data reduction step, the "signature" protein profiles that we obtain may not have a clear interpretation: a signature profile can be any pattern of protein intensity measurements and may not correspond to any protein intensity peaks.

To illustrate this first data analytic issue by a concrete example, consider the protein intensity measurements shown in the bottom panel of Figure 1. There are five large peaks (signals) that are visually evident in the plotted range of 7 500–10 000 mass/charge values. There are, however, other smaller peaks that are less evident as to whether or not they represent signals. A good mathematical definition of peaks would capture, at minimum, the five clear peaks, and possibly, less evident ones, but would also have a high level of specificity such that the rest of the data points would not be identified as peaks.

The second data analytic issue is the calibration of mass/charge measurements across multiple samples. This issue can be explained clearly by an example. Figure 2 shows the SELDI output from the first four subjects in the normal control group of the prostate cancer biomarker discovery project. In the left panel of Figure 2, it appears that, at least, the five visually apparent peaks, including the one marked by "×," and some less-evident peaks are aligned well in the direction of the mass/charge axis across the four samples. When we zoom into the small region near the peak marked by "×" (the right panel of Figure 2),

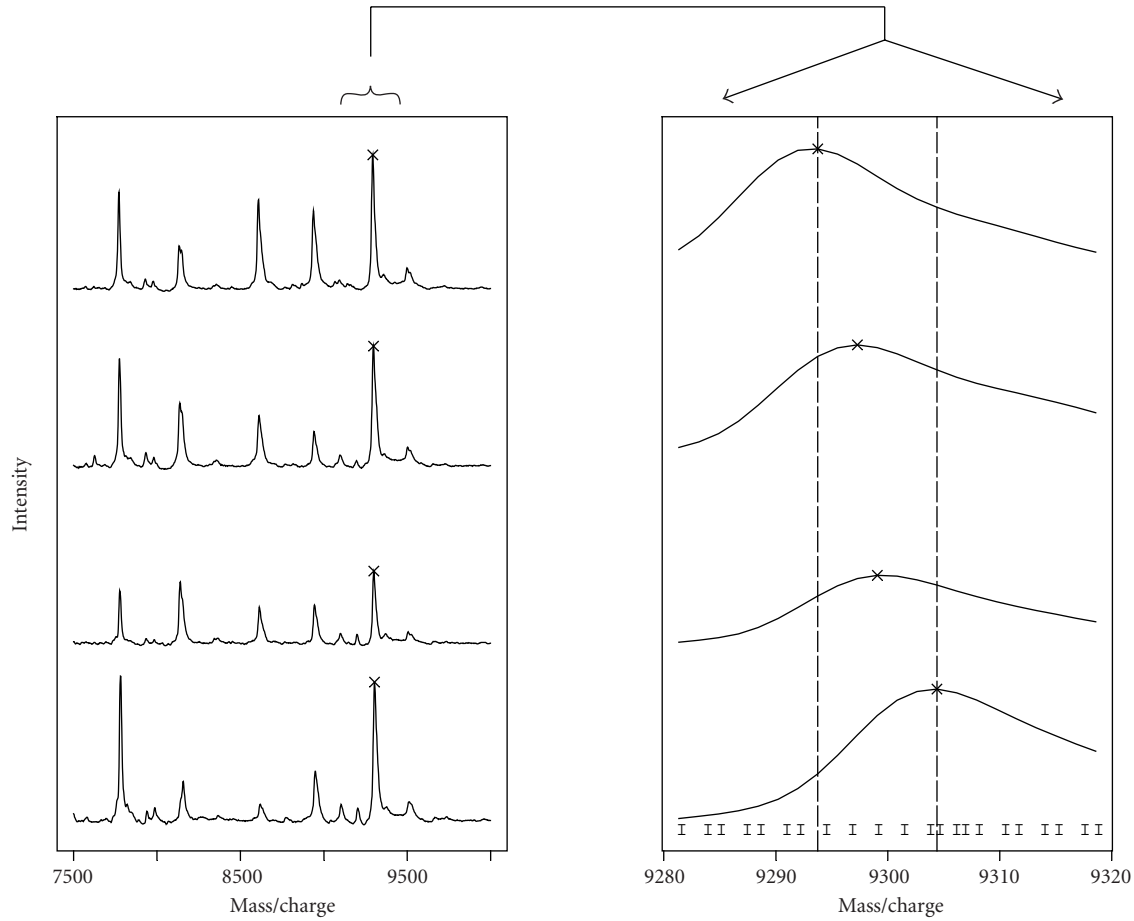


Figure 2. A set of SELDI output from four subjects: the left panel shows the protein intensity measures in the range of 7 500–10 000 mass/charge values and the right panel zooms into a small region corresponding to a visually apparent peak. The right panel shows slight shift in the mass/charge value of an identical peak across the subjects.

however, it becomes apparent that there is slight shift across the four samples with respect to the mass/charge values of the peak: there is an approximately 10-dalton shift between the first subject at the top and the fourth subject at the bottom. The measurement points in this small region are indicated by "|" at the bottom of the right panel of Figure 2: the 10-dalton shift corresponds to a 6-measurement-point shift in the mass/charge axis.

Although the magnitude of the shift is small, the inconsistent mass/charge values across samples for an identical peak present a challenge in data analyses: an identical peak is labeled by different mass/charge values across samples. Consider, in the right panel of Figure 2, the protein intensity at the mass/charge value of the peak from the first subject, that is, approximately at 9, 294 daltons. At this mass/charge value, the protein intensity for the second subject is nearly equal to the intensity value at the peak. For the third and fourth subjects, however, the protein intensities at the mass/charge value are approximately half and one quarter, respectively, of the intensities at the peaks. Thus, even though the magnitude of shifting is small, assessments of protein intensities by the original mass/charge values of the SELDI output would be greatly misleading.

The proposed solutions to the two data analytic issues

Our proposal for the first data analytic issue, the mathematical definition of "peaks," is to define peaks by judging, at each mass/charge point, whether or not the protein intensity at that point is the highest among its nearest $\pm N$-point neighborhood set, nearest in the mass/charge-axis direction; if it is the highest, that point is defined as a peak. We initially considered various values of N and chose $N = 10$ by trial and error in order to be on the inclusive side in classifying peaks; see also our previous discussion on the selection of N under a slightly different setting [6]. If a more conservative definition is preferred, the value of N can be increased.

Our proposal for the second data analytic issue, the calibration of mass/charge measurements across samples, is to replace the original mass/charge values of all peaks

with a set of calibrated mass/charge values. To describe our algorithm of the proposed method, we first introduce some terms/concepts that are helpful. We call a range of potential mass/charge shifting from a mass/charge point as the "*window of potential shift*" for that mass/charge point; and refer to the set of calibrated mass/charge values as the "*new mass/charge set*." Note that the window of potential shift for a mass/charge point, say P, contains the mass/charge point P itself. Based on quality-control experiments by the manufacturer of SELDI machines, it is known that the window of potential shift for a mass/charge point is approximately ±0.1–0.2% of the mass/charge value of that point; we used 0.2% in the current analysis.

The algorithm is initiated by calculating, at each mass/charge point, the total number of peaks, *in all samples*, that are within the window of potential shift for the mass/charge point. The mass/charge point that has the highest total number of peaks (summing over all samples) within its window of potential shift is entered into the new mass/charge set as a calibrated mass/charge value, and all the mass/charge points that are within the window of potential shift for this point are removed from the subsequent steps of the algorithm. Then, the above procedure is repeated (ie, finding the point, from the remaining points, that has the highest total number of peaks within its window of potential shift, entering it into the new mass/charge set, and removing the mass/charge points that are within its window of potential shift from the subsequent steps of the algorithm) until all peaks are exhausted from every sample.

The end product of this repeated procedure is the new mass/charge set. The final step of the algorithm is to construct a calibrated dataset that consists of protein intensity measures of each sample that correspond to the points in the new mass/charge set. For each sample, i, and for each point in the new mass/charge set, j, we propose to take the maximum protein intensity measure of the sample i, among the protein intensity measures corresponding to the window of potential shift for the point j, as the protein intensity measure Y_{ij} at the calibrated mass/charge point j. The final calibrated dataset is $\{Y_{ij}\}$ whose elements represent protein intensity measures indexed by the sample number i and the calibrated mass/charge value j.

An application of the calibrated dataset to biomarker discovery

To illustrate the utility of the calibrated dataset produced by the proposed method, we applied it to a construction of "signature profiles" of disease states in the prostate cancer biomarker discovery project. The 386 serum samples of the project were separated by a stratified random sampling into "test data" (a total of 60, 15 samples from each of the four disease states: late- and early-stage prostate cancer; BPH; and normal) and "training data" (the remaining 326 samples). The training data were used to construct a calibrated dataset, from which signature profiles were derived for classifying the three disease states of interest (ie, cancer, BPH, and normal). To test the performance of the derived signature profiles independently from the training data, the test data was used as follows. First, a calibrated *test* dataset $\{Z_{ij}\}$ was constructed by setting, for each test sample i and each calibrated mass/charge value j that appears in the derived signature profiles, the maximum protein intensity measure Y_{ij} within the window of potential shift for j as Z_{ij}. Second, the signature profiles derived from the training data were applied to the calibrated test dataset to classify each test sample into the three disease states. Finally, the classification errors in the test data were assessed comparing the classified disease states with the true disease states. The stratified sampling that created the training and test data was conducted by a statistician (DM) who received the data from the EVMS laboratory, and the disease states of the test samples were blinded to all data analysts.

In the analysis of the training data, the protein-intensity measure, Y_{ij}, for sample i at calibrated mass/charge value j, was transformed because of its heavily skewed distribution. The transformed protein intensity measure, T_{ij}, is given by

$$T_{ij} = \ln\left(Y_{ij} - c_j + 1\right) - s_i, \tag{1}$$

where c_j is the minimum protein intensity measure at the calibrated mass/charge value j among all training samples, which makes the logarithmic transformation possible, and s_i is the mean value of $\ln(Y_{ij} - c_j + 1)$ of the sample i across all calibrated mass/charge values. The subtraction of s_i aims to remove the sample-specific mean protein intensity since a number of sample-specific factors could modify the amounts and measurements of proteins across samples. The same transformation was also employed in the analysis of the test data $\{Z_{ij}\}$.

The signature profiles (classifiers of the disease states) were constructed using two logistic regression models: one for classifying cancer/BPH versus normal and the other classifying cancer versus BPH. The two logistic regression analyses used the respective disease states as outcome variables and transformed protein intensity measures $\{T_{ij}\}$ as potential covariates, selecting only those with significant associations with the disease states (at $P = .0001$ and $P = .0005$ levels, resp.) by a forward variable selection method.

RESULTS AND DISCUSSION

Results

Figure 3 shows the peaks identified by the proposed peak identification method. Our simple mathematical definition of peaks captured both visually apparent and some less-evident peaks. The number of peaks identified per sample was similar across the four disease states: the median (range) of 469 (361–571), 463 (389–596), 467 (367–555), and 444 (390–559) for the groups of late-stage

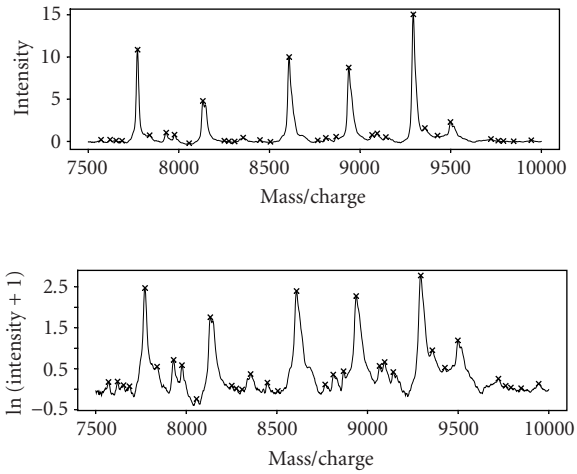


Figure 3. Peaks identified by the proposed peak identification method, marked by "×," in the range of 7, 500–10, 000 mass/charge values, corresponding to the bottom panel of Figure 1. The top panel shows the original protein intensity measures in the y-axis, while the bottom panel shows transformed protein intensity measures in the y-axis for ease in examining the peaks.

cancer, early-stage cancer, BPH, and normal control, respectively.

Figure 4 shows calibration of one of the visually apparent peaks in Figure 2 by the proposed method. The first four samples of the normal control group had peaks that were slightly shifted in the direction of the mass/charge axis around 9 300 mass/charge value. The calibrated mass/charge value corresponding to these peaks was 9 306.2 and, as shown in Figure 4, the four previously shifted peaks are now lined up at this calibrated mass/charge value with the original protein-intensity measures being used as $\{Y_{ij}\}$. The intensity values at 9, 306.2 mass/charge value changed from 7.87, 9.90, 6.40, 15.18 before the calibration to 8.61, 10.73, 7.04, 15.59 after the calibration in the four samples.

After the calibration, the number of mass/charge values in the new mass/charge set was 957. This represents a considerable reduction of data from 16 898 points per sample to 957 (5.7% of 16 898) in the range of 2 000–40 000 mass/charge values. Figure 5 shows the number of mass/charge values in the new mass/charge set according to their values. The number in the new mass/charge set was highest in the smallest values of mass/charge and monotonically decreased for larger mass/charge values.

The calibrated dataset was then used in the logistic regression analysis to construct signature profiles of three disease states (cancer, BPH, and normal). With the 957 log-transformed protein intensity measures as potential covariates, the forward selection method identified four calibrated mass/charge values that were significantly associated with the classification of cancer/BPH versus normal at $P = .0001$ level. Similarly, seven calibrated mass/charge values were selected into the model for the classification of

Table 1. Test data classification results of the logistic-model-based classifiers constructed using the calibrated training dataset.

Predicted by models	True disease state		
	Cancer	BPH	Normal
Cancer	27	2	0
BPH	1	13	0
Normal	2	0	15

cancer versus BPH at $P = .0005$ level. Note that the four and seven selected mass/charge values were those at which some samples showed peaks in protein intensity. Based on fitted probabilities from the two logistic regression models, the 60 test data samples were classified into the three disease states. Of the 60 test data samples, 55 (91.6%) were correctly classified, suggesting the high utility of the calibrated dataset, even with this simple classifier construction method using the standard forward selection logistic regression analysis.

Discussion

Previously in this project, the peak identification and mass/charge value calibration were performed manually, taking a significant amount of human effort and time [4]. The proposed data processing method was motivated to automate the human processing of the SELDI output by the use of computers. It mimics the steps of the previous manual processing and aims to eliminate potential human errors in dealing with the high-dimensional complex data. The excellent performance in classification shown in Table 1 suggests the high utility of the data produced by the proposed data processing method.

There are important advantages in separating the data processing stage, as proposed here, from the subsequent signature profile construction stage. First, the proposed method can be applied with complete blinding to the disease states of samples, ensuring an unbiased data processing across samples. Neither the peak identification nor the mass/charge-axis calibration of our proposed method requires knowledge of the disease state of each sample. Second, by separating the two stages that have distinct data analytic issues, we are able to consider various targeted approaches for resolving the stage-specific data analytic issues.

A limitation in our proposed method for the calibration of mass/charge values is that the calibration procedure depends on the dataset being analyzed, that is, the sets of calibrated mass/charge values from multiple datasets may not agree. This is perhaps not a critical issue at the biomarker discovery stage of research, such as the prostate cancer biomarker discovery project discussed here, since the data analysis for the biomarker discovery would use a single dataset. If the data are measured using multiple SELDI machines, either at one laboratory or at multiple laboratories, the dataset-dependent nature of

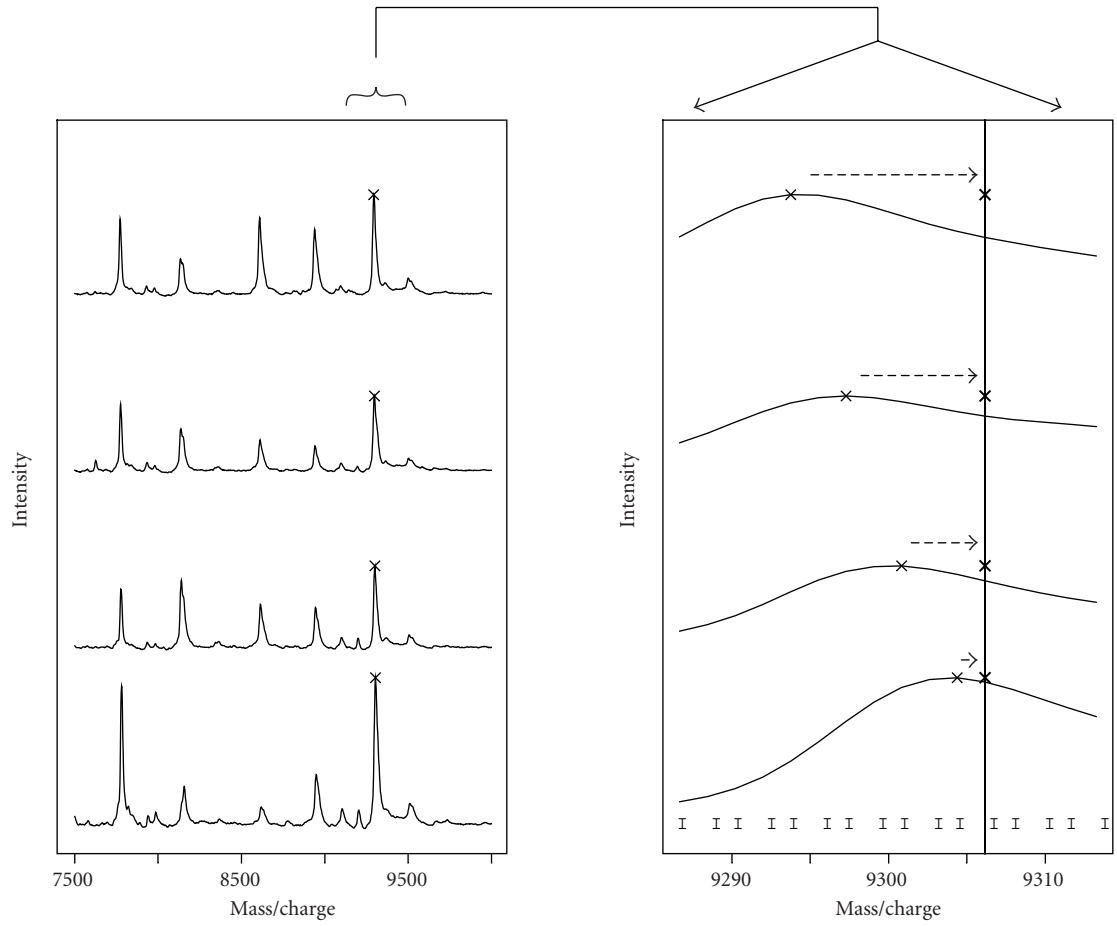


Figure 4. Calibration of a visually apparent peak in Figure 2 by the proposed method.

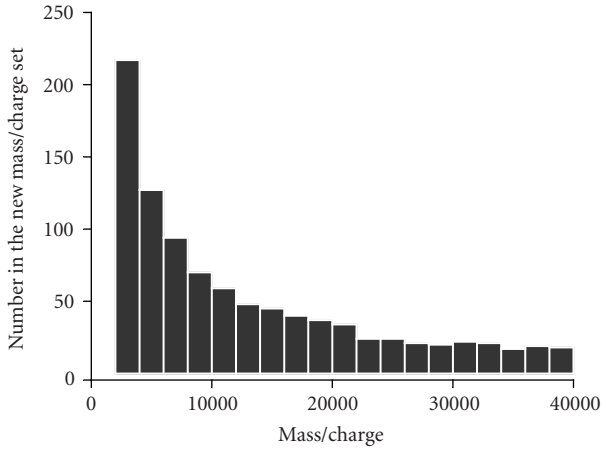


Figure 5. Number of mass/charge points in the new calibrated mass/charge set according to their mass/charge values.

the proposed method needs to be considered carefully. If datasets generated by multiple SELDI machines are combined, additional sources of variations are introduced (eg, between machine and laboratory variations). The "window of potential shift" of a mass/charge value in the combined dataset would, therefore, be expected to be larger than the ± 0.1–0.2% considered here, that was based on the quality-control data of the SELDI manufacturer. It is necessary to either use a wider window of potential shift or add an extra step in the method to minimize the additional variation when combining datasets generated by multiple SELDI machines.

A potential improvement of the proposed method is to make the mathematical definition of peaks such that it copies closely the thought process of experienced mass spectrometry experts in identifying peaks. Although our simple definition appears reasonable and functional, it would certainly enhance the method if a peak identification procedure similar to that of experts could be implemented. For example, a learning algorithm can be applied to a large dataset that is read and peak-identified by experts so that details of experts' procedures may be recognized for possible implementation in the mathematical definition.

In summary, we proposed an automatic data processing procedure for the peak identification and mass/charge-axis calibration problems for mass-spectrometer-

generated protein measures. Our procedure is easy to implement and appears to work effectively as evidenced in the excellent classification performance by the resulting calibrated dataset. There are points to improve in the proposed method and additional issues to resolve when it is applied to multiple datasets generated by more than one SELDI machine. We hope to resolve these issues in our future research.

ACKNOWLEDGMENT

This research was supported by Grant U01-CA86368 from the National Cancer Institute.

REFERENCES

[1] Srinivas PR, Srivastava S, Hanash S, Wright GL Jr. Proteomics in early detection of cancer. *Clin Chem*. 2001;47(10):1901–1911.

[2] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–577.

[3] 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.

[4] 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.

[5] Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem*. 2002;48(10):1835–1843.

[6] Yasui Y, Pepe M, Thompson ML, et al. A Data-Analytic Strategy for Protein-Biomarker Discovery: Profiling of High-Dimensional Proteomic Data for Cancer Detection. *Biostatistics* 2003;4(3):449–463.

[7] Srivastava S, Kramer BS. Early detection cancer research network. *Lab Invest*. 2000;80(8):1147–1148.

[8] Verma M, Wright GL Jr, Hanash SM, Gopal-Srivastava R, Srivastava S. Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. *Ann N Y Acad Sci*. 2001;945:103–115.

[9] Wright GL Jr, 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.

[10] Rubin RB, Merchant M. A rapid protein profiling system that speeds study of cancer and other diseases. *Am Clin Lab*. 2000;19(8):28–29.

[11] Adam BL, Vlahou A, Semmes OJ, Wright GL Jr. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*. 2001;1(10):1264–1270.

[12] Paweletz CP, Trock B, Pennanen M, et al. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis Markers*. 2001;17(4):301–307.

[13] Rosty C, Christa L, Kuzdzal S, et al. Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. *Cancer Res*. 2002;62(6):1868–1875.

* Corresponding author.
E-mail: yyasui@fhcrc.org
Fax: +1 206 667 5977; Tel: +1 206 667 4459

Journal of Biomedicine and Biotechnology • 2003:4 (2003) 249–255 • PII. S1110724303209244 • http://jbb.hindawi.com

RESEARCH ARTICLE

Selective Enrichment of Membrane Proteins by Partition Phase Separation for Proteomic Studies

M. Walid Qoronfleh,* Betsy Benton, Ray Ignacio, and Barbara Kaboord

Perbio Science, Bioresearch Division, 2202 N Bartlett Avenue, Milwaukee, WI 53202-1009, USA

Received 14 June 2002; accepted 18 December 2002

The human proteome project will demand faster, easier, and more reliable methods to isolate and purify protein targets. Membrane proteins are the most valuable group of proteins since they are the target for 70–80% of all drugs. Perbio Science has developed a protocol for the quick, easy, and reproducible isolation of integral membrane proteins from eukaryotic cells. This procedure utilizes a proprietary formulation to facilitate cell membrane disruption in a mild, nondenaturing environment and efficiently solubilizes membrane proteins. The technique utilizes a two-phase partitioning system that enables the class separation of hydrophobic and hydrophilic proteins. A variety of protein markers were used to investigate the partitioning efficiency of the membrane protein extraction reagents (Mem-PER) (Mem-PER is a registered trademark of Pierce Biotechnology, Inc) system. These included membrane proteins with one or more transmembrane spanning domains as well as peripheral and cytosolic proteins. Based on densitometry analyses of our Western blots, we obtained excellent solubilization of membrane proteins with less than 10% contamination of the hydrophobic fraction with hydrophilic proteins. Compared to other methodologies for membrane protein solubilization that use time-consuming protocols or expensive and cumbersome instrumentation, the Mem-PER reagents system for eukaryotic membrane protein extraction offers an easy, efficient, and reproducible method to isolate membrane proteins from mammalian and yeast cells.

INTRODUCTION

Based on the sequences from several genomes, transmembrane proteins have been predicted to comprise approximately 30% of eukaryotic proteomes [1]. Membrane proteins are the most elusive and the most sought after proteins in drug discovery. They play a key role in signal transduction, cell adhesion, and ion transport and are important pharmacological targets. Yet, because of their hydrophobic and basic nature, and frequently large size, their isolation is not easy. Traditional methods for membrane isolation are often cumbersome and protein yields are poor. Techniques used for membrane protein isolation include gradient separation [2], polymer partitioning [3], and chemical treatment [4]. These methods typically result in high purity but low recovery and, with the exception of polymer partitioning, are time consuming. Detergent extraction combined with ultracentrifugation is by far the most commonly used method for membrane protein isolation [5, 6, 7]; however, this method is a multistep process involving mechanical disruption of cells followed by lengthy centrifugation prior to solubilization of the proteins in detergent.

Nonionic detergents are widely used for the solubilization and characterization of integral membrane proteins. In particular, members of the Triton X series are commonly employed in phase separation of these proteins [6, 7]. We have developed a proprietary formulation and a protocol for the preparation of integral membrane proteins that is a nonmechanical alternative to traditional membrane protein isolation techniques. The protocol involves the gentle lysis of cells using a mild, proprietary detergent followed by membrane protein extraction utilizing the nonionic detergent, Triton X-114. Triton X-114 is a unique detergent in that it not only solubilizes membrane proteins but also separates them from hydrophilic proteins via phase partitioning at a physiological temperature [8]. Specifically, a solution of Triton X-114 is homogeneous at 0°C (forms a clear micellar solution) but separates into an aqueous phase and a detergent phase above 20°C (the cloud point) as micellar aggregates form and the solution turns turbid. With increased temperature, phase separation proceeds until two clear phases are formed where proteins partition according to their hydrophilic and hydrophobic features. Unlike traditional protocols involving phase partitioning with Triton X-114, our protocol does not require preparation of a membrane fraction as a prerequisite for protein solubilization. Membrane proteins are extracted directly from crude cell lysates quickly and efficiently with a standard benchtop microcentrifuge. The entire procedure is completed in one hour and has been optimized for the isolation of integral membrane proteins from a variety of mammalian cell lines as well as yeast cells.

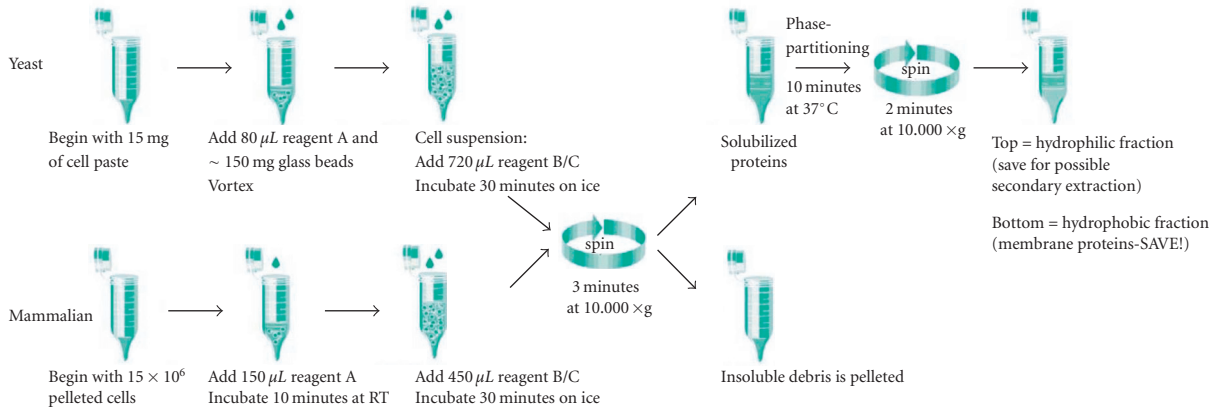


Figure 1. Schematic of Mem-PER Eukaryotic Membrane Protein Extraction Kit protocol. Each step of the procedure is outlined for a single extraction of either mammalian or yeast membrane proteins.

Detection and identification of proteins is facilitated through the enrichment of protein families and proteins in low abundance. Prefractionation of hydrophobic proteins enhances membrane proteomic analysis; therefore, it is essential to have reliable sample preparation methods that give high yields of this desired protein fraction. In this paper, we describe a fast, effective, and convenient protocol for membrane protein isolation involving temperature-induced phase separation of a proprietary formulation containing Triton X-114. We show that hydrophilic proteins (peripheral and cytosolic) are recovered in the aqueous phase whereas integral membrane proteins are enriched in the detergent phase. This procedure combines nonmechanical cell lysis with detergent fractionation/enrichment of membrane proteins and is termed the Mem-PER Eukaryotic Membrane Protein Extraction system.

MATERIALS AND METHODS

Cell culture conditions

Mammalian cell lines, rat brain C6, NIH-3T3, and HeLa, were obtained from American Type Culture Collection (Rockville, Md). The cells were grown to approximately 75% confluency in high-glucose Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS, 1% antibiotic/antimycotic agent (Gibco BRL), glutamine, and sodium pyruvate. Cells were incubated at 37°C in an atmosphere of 5% CO_2 and harvested with 0.25% trypsin. All cell culture reagents were obtained from HyClone, Inc (Logan, Utah) except where otherwise indicated.

Yeast culture conditions

Saccharomyces cerevisiae strain EGY-194 was grown in YPD media (MobiTec, Marco Island, Fla) at 30°C with agitation. Cells were harvested in the exponential growth phase at a density of 1–5 × 10^7 cells/ml with an A_{600} = 0.3–1.7.

Protein extraction protocol

The Mem-PER system consists of three reagents. Reagent A is a cell lysis buffer, reagent B is a detergent dilution buffer, and reagent C is a membrane solubilization buffer. A schematic of the protocol to extract and prepare either mammalian or yeast membrane protein fractions is shown in Figure 1. For yeast, approximately, 15 mg of wet yeast cell paste was resuspended in Mem-PER reagent A and vortexed with 150 mg of 405–600 micron acid-washed glass beads for 10 minutes to disrupt the yeast cell wall. The beads were collected by pulse centrifugation, and the cell suspension was transferred to a fresh tube. Subsequent fractionation was performed according to Figure 1. Cultured NIH-3T3, HeLa, and C6 cell lines were harvested using trypsin, and were washed and pelleted in phosphate-buffered saline (PBS) at 850 ×g for 2 minutes. Each cell pellet, containing 5×10^6 cells, was lysed at room temperature using Mem-PER reagent A. Yeast and mammalian membrane proteins were solubilized on ice with Mem-PER reagent C diluted 2 : 1 with Mem-PER reagent B. Reagents A and B/C were supplemented with Halt protease inhibitors cocktail (Pierce Biotechnology, Inc, Rockford, Ill). The solubilized protein mixture was centrifuged to remove cellular debris. The clarified supernatant was heated at 37°C for 10 minutes followed by centrifugation to produce separate membrane and hydrophilic protein fractions. Phase partitioning resulted in the hydrophilic proteins layering at the top and the hydrophobic membrane proteins layering at the bottom. A micropipette was used to carefully remove the top (hydrophilic) phase. The hydrophobic fraction was normalized to the volume of the hydrophilic fraction using Mem-PER reagent B diluted 4-fold with purified water. The fractions were further diluted 2-fold with diluted Mem-PER reagent B, to decrease the detergent concentration, and boiled in 6x-sample buffer. The isolated membrane protein fraction was used directly in SDS-PAGE and Western blotting.

Protein quantification

The Micro BCA Protein Assay Reagent Kit (Pierce) was used to quantify extracted membrane proteins. Mem-PER reagent C was initially found to interfere with the assay because it clouds at the required incubation temperature for the assay of 37°C; however, this interference was eliminated through dialysis using Slide-A-Lyzer MINI Dialysis Units (Pierce). Dialysis was performed overnight at 4°C against 25 mM Tris, pH 7.4, containing 0.5% CHAPS. CHAPS formed mixed micelles with reagent C, thereby raising the cloud point of the solution above 37°C. Approximately 100 μg of total protein was obtained from 5×10^6 C6 cells, and approximately 130 μg of total membrane protein was isolated from 15 mg of yeast cell pellet.

SDS-PAGE

Precast Novex brand (Invitrogen, Carlsbad, Calif) SDS-PAGE gels were utilized in all experiments. Standard electrophoresis conditions recommended by the gel manufacturer were employed.

Optional detergent removal prior to SDS-PAGE

Mem-PER reagent C was found to interfere with electrophoresis of low molecular weight proteins. Specifically, the detergent caused lane distortion and masked protein band visualization. This was remedied by treating Mem-PER-isolated membrane fractions with the PAGEprep Protein Cleanup and Enrichment Kit (Pierce). The kit contains a unique resin of modified diatomaceous earth that binds protein in an organic phase of dimethyl sulfoxide (DMSO), and allows contaminating chemicals and gel-incompatible material to be washed away. Cleanup was performed according to the manufacturer's instructions.

Antibodies

Mouse monoclonal antibodies against flotillin and acetylcholinesterase (AchE) were obtained from transduction laboratories (San Diego, Calif). Heat shock protein 90 (Hsp90), a polyclonal antibody raised in goats, and cytochrome oxidase subunit 4 (Cox4), a mouse monoclonal antibody, were obtained from Santa Cruz Biotechnology, Inc (Santa Cruz, Calif). For the yeast study, proteins were detected using monoclonal antibodies obtained from Molecular Probes, Inc (Eugene, Ore). Protein bands were visualized using antigoat and antimouse secondary antibodies conjugated with horseradish peroxidase from Pierce Biotechnology, Inc.

Western blotting and densitometric analysis

Protein fractions were prepared for electrophoresis by boiling in 6x-sample buffer. Prepared mammalian protein samples were separated using 4–20% Tris-glycine SDS-PAGE gradient gels while yeast protein samples were electrophoresed using NuPAGE 4–12% Bis-Tris gels. The protein fractions were then blotted to nitrocellulose. The blots were blocked in Superblock blocking buffer (Pierce) containing 0.05% Tween-20. After probing with primary and secondary antibodies, detection was performed with SuperSignal West Femto Maximum Sensitivity Substrate (Pierce) for 5 minutes followed by exposure to X-ray film for 15 seconds or to a FluorChem CCD camera (Alpha Innotech Corp, San Leandro, Calif) for 2 minutes. Bands were quantified using densitometry analysis (AlphaEaseFC software, Alpha Innotech) and expressed as a percentage of the total protein in the combined hydrophilic and hydrophobic (membrane protein) fractions.

RESULTS

Protein fractionation protocol

The membrane protein extraction protocol was accomplished in two parts (Figure 1). First, cells were lysed with a proprietary detergent and then a second proprietary formulation containing Triton X-114 was added to solubilize the membrane proteins. A white, flocculent material appeared following addition of the cell lysis component, Mem-PER reagent A. This debris was likely comprised of lipid and cell membrane material but not DNA since the addition of DNase was not found to diminish the particulate. Solubilization of the membrane proteins with Mem-PER reagent C diluted with Mem-PER reagent B was performed on ice with vortexing every 5 minutes. Longer incubation was not found to increase extraction efficiencies (data not shown). The cellular debris was removed during subsequent centrifugation. The hydrophobic proteins were then separated from the hydrophilic proteins through phase partitioning [8] at 37°C. Following careful separation of the two layers with a micropipette, membrane proteins were ready for subsequent analysis. Complete separation of the two layers was not possible due to the transient nature of the interface. No more than 10 samples were processed at one time since the interface slowly disappeared as the temperature of the sample fell below 37°C. Although a distinct separation could be seen, a small amount of crossover of each phase into the other could not be avoided during pipetting. In order to minimize contamination of the hydrophobic layer with the hydrophilic layer, some of the hydrophobic layer was sacrificed during removal of the top layer. A second round of extraction of the hydrophilic fraction obtained was not found to significantly increase membrane protein yields (data not shown). The hydrophilic, hydrophobic, and insoluble debris fractions were analyzed by SDS-PAGE and Western blotting.

Membrane protein extraction from mammalian cells

Mem-PER reagents were found to be highly efficient in the extraction of integral membrane proteins containing one or two transmembrane spanning domains. These results were found to be consistent with three different

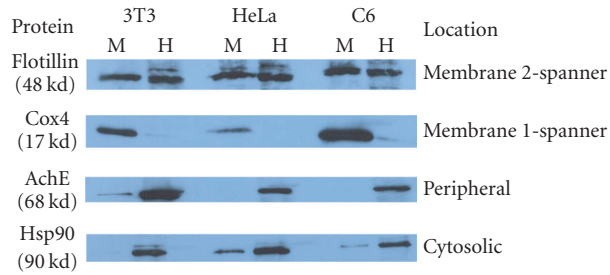


Figure 2. Partitioning of solubilized mammalian membrane proteins using the Mem-PER Kit. Proteins from three cell lines were solubilized and extracted using the Mem-PER Kit. Each set of hydrophilic and hydrophobic (membrane protein) fractions obtained was normalized to one another and analyzed by Western blot for four proteins from the cellular locations noted. PAGEprep resin was used to remove the detergent from the membrane fraction prior to SDS-PAGE/Western analysis of Cox4 due to the interference of detergent with band migration of low molecular weight proteins. A negligible amount of protein was found in all debris fractions (data not shown). Abbreviations: Acetylcholinesterase (AchE), cytochrome oxidase subunit 4 (Cox4), heat shock protein 90 (Hsp90), M = solubilized membrane protein fraction, H = hydrophilic protein fraction.

Table 1. Quantification of mammalian cell lysate proteins fractionated with the Mem-PER reagents in Figure 2.

Cell type	Fraction*	Flotillin (2-spanner)	Cox4 (1-spanner)	AchE (Peripheral)	Hsp90 (Cytosolic)
NIH-3T3	Membrane	45.0	90.8	1.7	6.8
	Hydrophilic	55.0	9.2	98.3	93.2
HeLa	Membrane	48.4	89.1	4.1	15.5
	Hydrophilic	51.6	10.8	95.9	84.4
C6	Membrane	56.0	94.5	6.4	10.6
	Hydrophilic	44.0	5.5	93.6	89.4

* Percent recovery of proteins following extraction is expressed as a percentage of the total protein in the combined hydrophilic and hydrophobic (membrane protein) fractions. The data was obtained from a single experiment but is representative of results obtained in multiple independent experiments.

mammalian cell lines, C6, NIH-3T3, and HeLa. As shown in Figure 2 and Table 1, the integral plasma membrane protein flotillin, containing two transmembrane domains, was extracted with an efficiency of approximately 50% from the three cell lines. These reported values were found to be reproducible in several isolated experiments. Extraction of Cox4, an outer mitochondrial membrane protein containing one transmembrane domain, was found to be even more efficient with approximately 90% recovery in the hydrophobic fractions obtained from the three cell lines. The membrane fraction probed for Cox4 (17 kd) in Figure 2 was treated with the PAGEprep resin prior to electrophoresis. The detergent in Mem-PER reagent C was found to interfere with electrophoresis of low molecular weight proteins (Figure 3) but did not affect electrophoresis of mid to high molecular weight proteins.

Cross-contamination of cytosolic and peripheral proteins into the prepared hydrophobic fraction was minimal. AchE, a peripheral protein, and Hsp90, a cytosolic protein, were routinely found to partition into the hydrophilic fraction with an efficiency of > 90%. The remaining 10% or less found in the hydrophobic fraction was likely due in part to difficulty in obtaining complete separation at the interface between the two phases, and the slow disappearance of the interface over time when the temperature fell below the cloud point of the mixture. Insoluble debris pellets typically contained < 5% of the membrane proteins examined in this study.

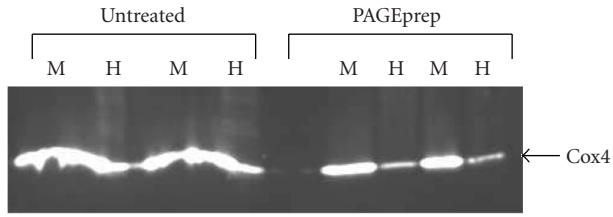


Figure 3. Removal of detergent from cell lysis fractions by the PAGEprep resin. Rat C6 cells were lysed and a membrane protein fraction isolated using the Mem-PER reagents. Membrane (M) and hydrophilic (H) cell fractions were separated by SDS-PAGE (4–20% gradient gel) with or without prior treatment using PAGEprep resin to remove detergent. Western blot analysis was performed as described in materials and methods using an antibody against Cox4. PAGEprep-treated samples show better band resolution than samples that were untreated and still contained the detergent.

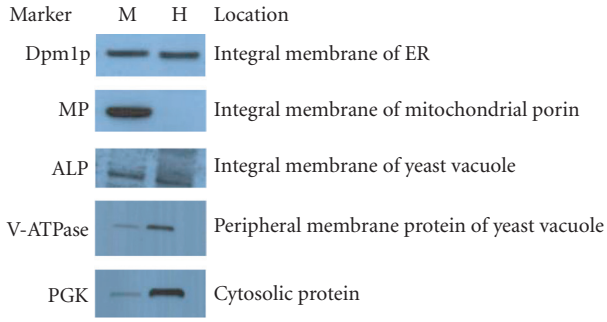


FIGURE 4. Partitioning of solubilized yeast membrane proteins using the Mem-PER Kit. Yeast proteins from *S cerevisiae* (strain: EGY-194) were solubilized and partitioned based on hydrophobic phase separation. Several proteins were solubilized and extracted using the Mem-PER reagents. Partitioning efficiency was determined through Western blot of normalized samples. A negligible amount of protein was found in all debris fractions (data not shown). Abbreviations: mitochondrial porin (MP), 3-phosphoglycerate kinase (PGK), alkaline phosphatase (AP), Dol-P-Man synthase (Dpm1p), M = solubilized membrane protein, H = hydrophilic protein fraction.

TABLE 2. Quantification of yeast proteins fractionated with the Mem-PER reagents in Figure 4.

Fraction*	Dpm1p	MP	AP	V-ATPase	PGK
Membrane	52.4	98.4	57.2	22.1	9.1
Hydrophilic	47.6	1.6	42.3	77.9	90.9

* Percent recovery of proteins following extraction is expressed as a percentage of the total protein in the combined hydrophilic and hydrophobic (membrane protein) fractions. The data was obtained from a single experiment but is representative of results obtained in three independent experiments.

Membrane protein preparation from yeast

Extraction efficiencies seen with the mammalian cell lines were similar to those obtained with yeast cells. Glass beads were used to lyse the rigid cell wall, and preparation of the membrane proteins was then performed with the Mem-PER reagents according to the same protocol used for the mammalian cells. Several protein markers were used to monitor the efficiency of the Mem-PER system to solubilize and isolate yeast integral membrane proteins. Figure 4 and Table 2 show the partitioning efficiency observed for these proteins. Mitochondrial porin (MP), an integral membrane protein of the outer mitochondrial membrane containing one transmembrane domain, was extracted into the hydrophobic fraction with an efficiency of greater than 90% with very negligible cross-contamination from the hydrophilic fraction. PGK (3-phosphoglycerate kinase), a cytosolic protein, was extracted into the hydrophilic fraction with an efficiency of greater than 85%. Alkaline phosphatase (ALP), an integral membrane protein of yeast vacuoles, was extracted into the hydrophobic fraction with an efficiency of > 50% whereas greater than 70% of V-ATPase, a peripheral membrane protein of yeast vacuoles, was recovered in the hydrophilic fraction. Dol-P-Man synthase (Dpm1p), a membrane protein in the yeast endoplasmic reticulum containing one transmembrane spanning domain, was extracted into the hydrophobic fraction with an efficiency of 50%. Similar results were obtained for all of these proteins in several isolated experiments. Insoluble debris pellets typically contained between 3 and 20% of the yeast membrane proteins examined in this study.

DISCUSSION

Transmembrane proteins are a valuable family of proteins. Functionally they are central to cell life and are the target of about 80% of all drugs. Preparation of membrane proteins is time consuming and difficult; therefore, development of analytical systems that allow the isolation and identification of this group of proteins would be desirable. Ideally, the isolation process should be mild yet rapid. Detergents have played significant roles in this effort [6, 7]. Detergents serve as invaluable tools to isolate, solubilize, and manipulate membrane proteins for subsequent biochemical and biophysical characterization [9]. Consequently, our understanding of the structure and function of membrane proteins has advanced significantly over the past decade. Nonionic detergents have been useful in this regard since they are widely used for the solubilization and characterization of integral membrane proteins. These proteins can be separated from hydrophilic proteins using the nonionic detergent Triton X-114 that undergoes separation at physiological temperatures into detergent-rich and aqueous-rich phases [8].

Many fractionation protocols exist for the enrichment of hydrophobic proteins [4, 6]; however, isolation of these proteins can be a tedious and time-consuming process requiring gradient methods and expensive ultracentrifugation equipment. A more convenient fractionation of membrane proteins can be achieved through the use of detergents. We have developed a proprietary mild detergent formulation and a protocol for the lysis of cells followed by enrichment of hydrophobic proteins via phase separation. Our unique cocktail contains Triton X-114. The protocol was performed with a benchtop microcentrifuge and did not require mechanical lysis such as sonication or Dounce homogenization. Separations were performed on a microscale; however, similar methodology using phase partitioning has been demonstrated on a large preparative scale [10]. We obtained membrane protein extracts quickly and efficiently from mammalian cells. Yeast membrane proteins were obtained in a similar fashion, except that the yeast cell wall was first removed. Glass beads were found to disrupt the cell wall quickly and efficiently [11] and were effective when used in combination with the cell lysis reagent, Mem-PER reagent A. Extraction efficiencies of approximately 50% or greater were typically seen with proteins containing one or two transmembrane spanning domains. Lower yields may be obtained with more complex integral membrane proteins, and variability in extraction efficiency may be observed depending on factors such as posttranslational modifications, and the number of transmembrane spanning domains. In addition, anomalous partitioning of some integral membrane glycoproteins has been observed. For example, intact acetylcholine receptor, an integral membrane protein containing four transmembrane domains, has been shown to partition into the aqueous-rich hydrophilic phase [12]. The reason for this behavior may be due to the large hydrophilic moieties on the glycoprotein and/or the channel-forming property of this protein. Membrane protein activity may be maintained following separation into the mild detergent environment of the Mem-PER system. Several integral membrane proteins have been found to retain their biological activity when solubilized in nonionic detergents [13]; however, retention of activity is dependent on the characteristics of the protein being analyzed and cannot be assured.

Triton X-114 is an effective reagent for the isolation of membrane proteins from mammalian systems. This method of fractionation has been used to isolate 75% of the integral membrane glycoproteins from prepared erythrocyte membranes [8] and nearly 100% of cytochrome b558 from prepared bovine granulocytes [14]. It has also been used to solubilize membrane proteins of subcellular fractions from the bovine adrenal medulla [15] as well as hepatic Golgi membrane proteins [16]. In all of these reports, an initial purification was performed prior to phase partitioning in Triton X-114. Our protocol was designed for crude cell lysates and does not require prior processing.

Membrane protein extracts obtained using Triton X-114 have been used in many downstream applications. Golgi proteins partitioned with the nonionic detergent were analyzed by mass spectrometry following one-dimensional SDS-PAGE [16]. Hydrophobic proteins isolated using Triton X-114 have also been analyzed by 2D gel electrophoresis following removal of the detergent using hydroxyapatite column chromatography [17]. A combination of 2D electrophoresis and mass spectrometry was used to identify a hydrophobic receptor protein, very-low-density lipoprotein, in the detergent-enriched phase and the cytoplasmic protein Hsp90 in the aqueous phase. Clearly, phase separation of Triton X-114 is a useful tool in the prefractionation of membrane proteins, and the detergent-rich extracts obtained with this method have been used successfully in proteomics applications.

The Mem-PER system is an excellent tool for the initial purification and preparation of protein fractions for downstream analysis. It provides a rapid and convenient protocol for the reproducible partitioning of mammalian and yeast proteins into hydrophobic and hydrophilic fractions. Interestingly, the extraction of yeast membrane proteins has never been performed with phase partitioning, and to our knowledge this is the first report. Prefractionation of complex protein mixtures is critical for proteomic studies because it increases the resolving power of many analytical techniques by allowing for the identification of low-abundance proteins. Selective separation of hydrophobic proteins enhances membrane proteomic examination. Integral membrane proteins cannot be extracted easily; however, phase partitioning is a proven and valuable technique for the enrichment of this important protein family.

REFERENCES

[1] Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*. 1998;7(4):1029–1038.

[2] Morre DJ. Isolation and purification of organelles and endomembrane components from rat liver. In: Chrispeels MJ, ed. *Molecular Techniques and Approaches in Developmental Biology*. New York, NY: John Wiley;1973:1–27.

[3] Morre DJ, Morre DM. Preparation of mammalian plasma membranes by aqueous two-phase partitioning. *Biotechniques*. 1989;7(9):946–958.

[4] Lenstra JA, Bloemendal H. Topography of the total protein population from cultured cells upon fractionation by chemical extractions. *Eur J Biochem*. 1983;135(3):413–423.

[5] Ohlendieck K. Extraction of membrane proteins. In: *Protein Purification Protocols*. *Methods in Molecular Biology*. vol. 59. Totowa, NJ: Humana Press Inc;1996:293–304.

[6] Helenius A, Simons K. Solubilization of membranes

by detergents. *Biochim Biophys Acta*. 1975;415(1): 29–79.
[7] Tanford C, Reynolds JA. Characterization of membrane proteins in detergent solutions. *Biochim Biophys Acta*. 1976;457(2):133–170.
[8] Bordier C. Phase separation of integral membrane proteins in Triton X-114 solution. *J Biol Chem*. 1981;256(4):1604–1607.
[9] Garavito RM, Ferguson-Miller S. Detergents as tools in membrane biochemistry. *J Biol Chem*. 2001;276(35):32403–32406.
[10] Pryde JG. Triton X-114: a detergent that has come in from the cold. *Trends Biochem Sci*. 1986;11:160–163.
[11] Panaretou B, Piper P. Isolation of yeast plasma membranes. In: Evans IH, ed. *Yeast Protocols: Methods in Molecular Biology*. vol. 53. Totowa, NJ: Humana Press Inc;1996:117–121.
[12] Maher PA, Singer SJ. Anomalous interaction of the acetylcholine receptor protein with the nonionic detergent Triton X-114. *Proc Natl Acad Sci U S A*. 1985;82(4):958–962.
[13] Tzagoloff A, Penefsky HS. Extraction and purification of lipoprotein complexes from membranes. *Methods Enzymol*. 1971;XXII:204–219.
[14] Pember SO, Heyl BL, Kinkade JM Jr, Lambeth JD. Cytochrome b558 from (bovine) granulocytes. Partial purification from Triton X-114 extracts and properties of the isolated cytochrome. *J Biol Chem*. 1984;259(16):10590–10595.
[15] Pryde JG, Phillips JH. Fractionation of membrane proteins by temperature-induced phase separation in Triton X-114. Application to subcellular fractions of the adrenal medulla. *Biochem J*. 1986;233(2):525–533.
[16] Bell AW, Ward MA, Blackstock WP, et al. Proteomics characterization of abundant Golgi membrane proteins. *J Biol Chem*. 2001;276(7):5152–5165.
[17] Wissing J, Heim S, Flohe L, Bilitewski U, Frank R. Enrichment of hydrophobic proteins via Triton X-114 phase partitioning and hydroxyapatite column chromatography for mass spectrometry. *Electrophoresis*. 2000;21(13):2589–2593.

* Corresponding author.
E-mail: walid.qoronfleh@perbio.com
Fax: +1 414 227 3759; Tel: +1 414 227 3605