Informatic Tools for Proteome Profiling

Deb N. Chakravarti¹, Bulbul Chakravarti², and Ioannis Moutsatsos³

¹Biotechnology Discovery Research Group, Wyeth Lederle Vaccines, West Henrietta, NY; ²Department of Chemistry, Rochester Institute of Technology, Rochester, NY; ³Proteome Informatics Group, Department of Genomics, Genetics Institute, Wyeth-Ayerst Research, Cambridge, MA, USA

Computational Proteomics Supplement 32:S4-S15 (March 2002)

ABSTRACT

In recent years, the practice of proteomics research has experienced a dramatic shift within the pharmaceutical and biotechnology industry with the widespread implementation of novel applications. The areas of interest extend all the way from discovery of novel drug, vaccine, and diagnostic targets, characterization of protein-based products, toxicology, and identification of surrogate markers of activity in clinical research, to the ability to provide information on the mechanisms of drug action. The power of two-dimensional gel electrophoresis as well as advances in mass spectrometric techniques combined with sequence database correlation have enabled speed and accuracy in identification of proteins in complex mixtures. This article surveys currently available software and informatic tools related to these methods for proteome profiling. The broad acceptance of these technologies, however, has not been accompanied by significant advances in the informatics and software tools necessary to support the analysis and management of the massive amounts of data generated in the process. In this context, this article also discusses the importance of relational databases for protein identification data management.

INTRODUCTION

The term “proteome” has been defined as the PROTEins expressed by the genOME of an organism or tissue. The analysis of the proteome—or proteomics—involves identification as well as qualitative and quantitative comparison of proteins expressed under different conditions, elucidation of their properties such as covalent structure, tertiary structure, intermolecular interaction, and functions usually in a large-scale, high-throughput format. In contradistinction to large scale, the hallmarks of high throughput are real-time data analysis and the absence of human intervention for verification of data. Progress in proteomics has been possible because of advancements in several different areas within the last ten years. These include advances in protein analytical methods using mass spectrometry, the availability of the complete genomic sequence of a number of organisms, the development of protein/peptide microarrays, computational methods, and computer technology. For proteome profiling of a particular system or organism, a number of specialized software tools are necessary. A vast amount of information that needs to be evaluated critically is generated within a short period of time in any proteomics project. To meet these

demands, the discipline of proteome informatics is rapidly growing as a specialized branch of bioinformatics. The latter has been defined as the field of science in which biology, computer science, and information technology merge into a single discipline (44). Proteome informatics involves collection, storage, search, analysis, classification, integration, management, and retrieval of data in conveniently accessible databases. Strong proteome informatics support is indispensable in proteomics projects to obtain seamless and transparent access to accurate information from multiple sources in a platform as well as database structure-independent manner (10). As has been suggested about bioinformatics (50), the real long-term value of proteome informatics is believed to lie not so much in the tools but in the conversion of the knowledge that proteome informatics would deliver into better prophylactic, therapeutic, and diagnostic products.

The proteome profile of an organism or a tissue is like a snapshot containing the qualitative and/or quantitative protein profile under a specific set of conditions. It is a dynamic entity, as opposed to the genome that can be considered fixed. The information generated in proteomics studies includes protein expression levels, co-translational and post-translational modification, subcellular localization, signaling pathways, protein complexes, protein-protein interactions, and the like. The enormity of data processing involved in an average-sized proteome project is mind-boggling. In this context, it is particularly interesting to note that plans to elucidate the human proteome have been discussed at a meeting organized by the recently created Human Proteome Organization or HUPO (www.hupo.org) (1). However, in contrast to the Human Genome Project, which, in general, used the basic technology for gene sequencing, proteomics involves application of different novel technologies, many of which are still in the process of development or standardization. The use of peptide mass spectrometric data to search sequence databases for identifying proteins is now rather well-established. The technique has recently been applied to search large genomes like the human genome (12,38). It is becoming possible to apply peptide mass spectrometric data to obtain experimental confirmation or verification of predicted coding sequences, identification of novel genes and elucidation of splicing patterns (11).

At present experimental technologies used for large-scale and/or high-throughput analysis of proteomes directed at protein identification and their qualitative and quantitative com-

parison can be classified into three different categories:

1. Two-dimensional (2-D) gel electrophoresis-based methods followed by mass spectrometry.
2. Direct analysis of complex mixtures of proteins using mass spectrometry.
3. Analysis of a complex mixture of proteins using microarrays of antibodies, peptides, etc. (9,29).

In this article, we will focus on the software and informatic tools related to the first two approaches. These tools used for other proteomic approaches, such as protein microarrays, biomolecular interaction analysis, yeast two-hybrid projects, high-throughput structural proteomics, protein-protein interaction (protein interactome) pathways and cell signaling networks, elucidation of protein function through large-scale computational analysis of sequence and structure are outside the scope of this article.

INFORMATIC AND SOFTWARE TOOLS FOR THE SAMPLE FRACTIONATION OR SAMPLE PREPARATION STAGE

For 2-D gel electrophoresis-based proteomics methods, the initial step involves fractionation of the sample by 2-D gel electrophoresis. The separated protein spots are excised, enzymatically digested with a site-specific protease such as trypsin followed by mass spectrometric analysis and sequence database searching for protein identification. The main software tool necessary for the sample fractionation/sample preparation stage of this approach is that for 2-D gel image analysis, which will be discussed below. The software tools used to operate 2-D gel spot excision robots are instrument-specific and are not covered in this article. In large proteomics projects, the information on samples used in 2-D gel electrophoresis is often recorded as part of a laboratory information management system (LIMS). The alternative approach to 2-D gel analysis usually involves direct digestion of the protein or protein mixture with a site-specific protease such as trypsin followed sequentially by liquid chromatography (LC) or capillary electrophoresis (CE) coupled with on-line tandem mass spectrometry (MS/MS). The operational software for LC and CE are instrument-specific and are not covered in this article. The samples used in this approach may have been pre-fractionated by immuno-precipitation, affinity purification, and/or sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

2-D Gel Image Analysis Software

2-D gel image analysis software has been in development (2,23,40) since the introduction of this electrophoretic technique in the mid-1970s, and further improvements were made in the late 1980s (4,5,22,51,53). A list of some of the commercially available 2-D gel image analysis software available now is presented in Table 1. All of these packages have been designed to digitize images from 2-D gels, detect, quantify, and identify spots of differential intensity between gels by matching the control and the experimental gel(s), and generate reports listing the differentially expressed protein spots. However, individual packages may apparently differ in several aspects, such as the com-

puter platform they can run on, the file format used, or the graphical user interface. Despite some limitations of the 2-D gel electrophoretic technique, it is currently the most widely used approach for obtaining global quantitative information about the expression level of all proteins in a sample that can be visualized on the gel using appropriate technology (52). Protein identification using mass spectroscopy is typically performed as the step next to 2-D gel electrophoresis. Some of the 2-D gel image analysis packages are able to interact with automated robotic systems to direct the excision of spots of interest for subsequent mass spectrometry analysis.

For setting up a commercial or high-throughput environment for proteome analysis, it is essential that the 2-D gel image analysis software support robust database tools for storing images, as well as data from spot analysis, quantification, and identification. Furthermore, these data need to be viewed and frequently clustered in the context of the experimental studies in which they were performed. However, many of these 2-D gel image analysis software packages lack robust integration with the powerful relational databases that currently store most enterprise data. The various proprietary storage formats used by the commercial vendors further complicate the accessibility and on-line availability of useful information generated by the image analysis software. However, the scenario is changing; for example, Scimagix (San Mateo, CA, USA) has developed its 2-D gel image analysis product, ProteinMine™, on the commercial Oracle® relational database system (Oracle Corporation, Redwood Shores, CA, USA) used by the company for their scientific image management system (SIMS). This kind of open storage architecture may be of significant value to laboratories seeking to integrate with other LIMS packages they are using.

INFORMATIC AND SOFTWARE TOOLS FOR THE PROTEIN IDENTIFICATION STAGE

Identification of Proteins by Mass Spectrometry

In a proteomics project, mass spectrometry has become the core technology for protein identification by matching peptide mass spectra to a sequence database (7,60). For this, a protein or a mixture of proteins is digested with a proteolytic enzyme, usually trypsin, and the peptides analyzed using appropriate mass spectrometric techniques. The mass spectrometric peak extraction software that converts the mass spectrometric peaks into a list of mass values for submission to the sequence database search engines are not covered in this article. As it stands now, there are three different ways (described below) to use mass spectrometric data to identify proteins by searching sequence databases. The first of these involves determination of the molecular weights of the peptides in the unfractionated digest by mass spectrometry (MS). For the other two approaches, tandem mass spectrometry (MS/MS) is carried out, which provides amino acid sequence-specific fragmentation patterns of individual peptides derived from the protein. In MS/MS experiments, initial measurement of the mass of a peptide is followed by its fragmentation and re-analysis of fragment mass spectra. The pattern of fragment ions produced is fairly unique for a given peptide sequence. The value of the MS/MS spectra lies in the

Table 1. Some Commercially Available 2-D Gel Image Analysis Software

Name of Software	Source
DeCyder™ 2D Analysis	Amersham Biosciences (www.apbiotech.com)
Delta2D™ (www.decodon.com/Solutions/Delta2D.html)	DECODON GmbH (www.decodon.com)
GELLAB II+ (www.scanalytics.com/product/gel/Gellab.html)	Scanalytics (www.scanalytics.com)
GD Impressionist™ (www.genedata.com/products/impressionist)	GeneData (www.genedata.com)
ImageMaster™ 2D Elite	Amersham Biosciences (www.apbiotech.com)
ImageIQ™ (www.proteomesystems.com/product/profile.asp?Category=gel+image+analysis)	Proteome Systems Ltd. (www.proteomesystems.com)
Investigator™ HT PC Analyzer* (www.genomic solutions.com/proteomics/2dgelanal.html)	Genomic Solutions (www.genomicsolutions.com)
Melanie 3 (www.genebio.com/Melanie.html)	GeneBio (www.genebio.com)
PDQuest™ 2-D Analysis Software	Bio-Rad Laboratories (www.bio-rad.com)
Phoretix™ 2D (www.phoretix.com/products/2d_products.htm)	Nonlinear Dynamics (www.nonlinear.com)
Progenesis™ (www.nonlinear.com/2D/progenesis)	Nonlinear Dynamics (www.nonlinear.com)
ProteinMine™ (www.scimagix.com/products-ProteinMine.html)	Scimagix (www.scimagix.com)
ProteomeWeaver (www.definiens-imaging.com/proteomeweaver)	Definiens Imaging GmbH (www.defeniens-imaging.com)
TotalLab (www.totallab.com)	Nonlinear Dynamics (www.nonlinear.com)
Z3 2D-Gel Analysis System (www.2dgels.com)	Compugen Ltd. (www.compugen.co.il)
*Developed in collaboration with Nonlinear Dynamics.	

fact that the dissociation of the precursor peptide ion occurs along known pathways. For fragmentation of protonated linear peptides, the accepted nomenclature uses a, b, c to indicate common types of product ions that contain the N terminus and x, y, z for those ions that contain the C terminus. The difference in mass between two consecutive product ions of the same letter type represents the mass of the intervening amino acid residue. Information from internal fragments that contain neither terminus can also help establish or confirm peptide composition. The different methods to identify proteins based on mass spectrometric data are as follows:

1. *Peptide Mass Fingerprint or Peptide Mass Map Analysis.* A set of peptide molecular weights obtained from a proteolytic digest of a protein represents a unique fingerprint. Proteins are identified based on comparisons of peptide molecular weights determined by mass spectrometry, with the theoretical masses of peptides produced in silico by digestion of sequences in a target database (proteins or translated nucleotides) with the same enzyme (13,31,36,42, 46,49, 61, 62). A review of the earlier literature on this subject with special reference to database searching is available (15).
2. *Peptide Sequence or Peptide Sequence Tag Query.* In this approach, peptide MS/MS data are combined with amino acid sequence and/or composition data to identify the protein from a protein (or nucleotide) sequence database. Usually, a partial sequence of the peptide is manually generated by subtracting the mass values of suitable fragment ions. This information together with the mass data allows for more specific probing of sequence databases (13,17,20,41,49).

3. *MS/MS Ion Search Analysis.* Uninterpreted MS/MS data from a single peptide or from a series of peptides in a complete LC-MS/MS run are matched with protein sequences in a protein (or nucleotide) sequence database to identify proteins without any manual sequence interpretation (17,49). The input data is usually a list of fragment ion mass and intensity values.

For samples containing only a single protein or a mixture of a few proteins, the peptide mass fingerprint approach is suitable. This can be conveniently performed in a fast high-throughput manner using a matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectrometer. Increased mass accuracy using delayed extraction usually allows for a better distinction of true hits from false positives in database correlation searches for protein identification. Using automated systems suitable for high-throughput analysis currently available, a protein digest can be analyzed in a MALDI-TOF mass spectrometer and the protein identified simultaneously (i.e., in real time) by searching a sequence database within a minute. For identification of the components of a simple mixture of proteins using the peptide mass fingerprint approach, it is necessary to determine the peptide mass with high accuracy (37,62). For analysis of more complicated protein mixtures, or for confirmation of results obtained by peptide mass fingerprint analysis, the peptide sequence tag query or the MS/MS ion search approach is used. In terms of protein identification, both of these offer a greater degree of specificity than the peptide mass fingerprint approach. However, in terms of speed, both of these methods are much slower than the peptide mass fingerprint approach, and, in addi-

tion, the sequence tag query requires manual interpretation of data. In terms of real time for database searches, the sequence tag approach is faster than MS/MS ion search because the former is just a filter on the database. For “difficult” MS/MS spectra, that is, spectra with a high level of noise, absence of fragment ion ladders, etc., the MS/MS ion search is the method of choice.

Tandem mass spectrometry is usually performed in a triple-quadrupole, quadrupole ion trap, or quadrupole time-of-flight hybrid (Q-TOF) mass spectrometer often with on-line LC capability. A different type of instrument, the Fourier transformation cyclotron resonance mass spectrometer (FTICR-MS) that offers considerably increased sensitivity, resolution, and accuracy for mass measurement, has enormous potential for use in proteomics. It has been demonstrated that with added constraints, the highly accurate mass of peptides (“accurate mass tags”) obtained by peptide mass fingerprint analysis in this type of instrument is sufficient for protein identification from correlative database searching without the need for time-consuming tandem mass spectrometric analysis (14,26).

Several software tools for database searches using mass spectrometric data for protein identification are available (Table 2). Details about many of these and issues involved in interpretation of the results have been reviewed (7,19). Although, many of these search engines can be used freely over the Internet, it is often necessary to license these softwares for in-house use. The reasons for doing so derive from specific requirements, such as confidentiality issues, ability to search custom or proprietary sequence databases, need for automated or high-throughput work, or customized import of data into a database for subsequent mining. Among the tools described in Table 2, at least Mascot and Sequest can be run on computer clusters. Although the details regarding the computer clusters and computer farms in large-scale industry-based proteomics projects are often proprietary, some information is available on their scope and magnitude. For example, it has been reported that GeneProt’s large-scale proteomic discovery center in Geneva, Switzerland, will use 1420 Compaq™ Alpha-based Tru64 UNIX computer processors to capture, store, and analyze the terabytes of proteomics data that will be generated by 51 mass spectrometers (24).

Practical Considerations for Protein Identification from Mass Spectrometric Data

Database search methods for identification of proteins from mass spectrometric data usually give a score representing the match of the experimental data with the theoretical data based on the identified sequence in the database. The score is simply the number of matches in some algorithms, whereas in other algorithms some other criteria are used in addition. For example, the simplest method for peptide mass fingerprint analysis is based on ranking of the proteins in the database according to the number of peptides matching with the experimentally determined peptide mass values. Protein sequences are arranged according to the score and presented in a report format. For a pure protein band on a gel, the best score often stands out from the rest of the proteins in the list and is selected. On the other hand, when the sample contains a mixture of proteins, each database hit from MS/MS data, particularly those scoring below an arbitrary cut-off value, determined by the method used, often needs

to be independently validated. It is a difficult and time-consuming task to manually sift through large amounts of data containing multiple matches to identify spectra that can be ignored because of poor signal or contaminants and to detect incorrect database matches that can be considered as false positives. The simplest validation for proteins identified from MS/MS data is to check if several different peptides are found to match to the same protein. However, for a complex mixture of proteins, quite often a few or a single peptide from a particular protein can be used for identification of the protein. In addition, independent of the type of scoring system used, there is a risk of obtaining a false identification due to random matching of peptide masses. The probability of a false identification can be calculated if the score-frequency for random identifications is known (7). One approach for obtaining the latter is by computer simulations (18). An alternative approach involves direct calculation of the probability that a certain protein sequence matches the experimental data (8,49). For high-throughput methods, where the human scrutiny of the data is not practicable, probability-based scoring schemes are preferred so that standard statistical confidence tests can be applied. Using computer simulation, a study on the application of chemical derivatization techniques for improving the efficiency of protein identification from mass spectrometric data has been described (54).

In contradistinction to automated correlative database searching, it is possible to determine the amino acid sequence of peptides *de novo* from the MS/MS data, and this typically involves manual interpretation by expert mass spectrometrists. *De novo* sequence analysis of peptides in an automated format remains an open problem in mass spectrometry. However, algorithms being developed for such purposes can be used to validate the results obtained from correlative database search in an automated, high-throughput format. One such algorithm is SHERENGA (16), which can generate a ranked list of potential peptide sequences from the MS/MS spectra. A similar approach (56) is based on the premise that if the determined *de novo* sequences (which may be correct or incorrect) can account for the MS/MS data better than a database-derived sequence, then the sequence obtained from the database search is likely to be a false positive. On the other hand, if the database search-derived sequence accounts for the MS/MS data better than or equal to the *de novo* sequences, then it is likely to be a proper match. In another approach, the ability of a mass spectrometric method, using per-methyl esterification of peptides for relative quantification of proteins between two mixtures of proteins, for automated *de novo* sequence derivation from the same dataset has been described (25).

In a proteomics project aimed at identification of proteins, it is not uncommon to find that mass spectrometric data cannot be correlated with any sequence in the database searched. Several possible reasons for this are discussed below.

1. *Absence of the sequence in the database searched.* The peptide(s) may be derived from novel proteins or from allelic or strain- or species-derived variants of known proteins. It is also possible that there may be errors in the reported sequence. If necessary, *de novo* sequence analysis of peptides from the MS/MS spectra may be used to determine related proteins using homology-based database search methods. A homology-based sequence database search program, CIDentity (57), which is a modified version of FASTA

Table 2. Some Database Search Tools for Protein Identification Based on Mass Spectrometric Data

Name of Software	Source
Peptide Mass Fingerprint or Peptide Mass Map Analysis	
Mascot ^a (www.matrixscience.com)	Matrix Science Ltd.
MassSearch ^a (cbrg.inf.ethz.ch/Server/MassSearch.html)	ETH
Mowse ^a (www.hgmp.mrc.ac.uk/Bioinformatics/Webapp/mowse)	UK Human Genome Mapping Project Resource Centre
MS-FIT ^a (prospector.ucsf.edu)	UCSF
PepMAPPER ^a (wolf.bms.umist.ac.uk/mapper)	UMIST
PepSea ^a (195.41.108.38/PepSeaIntro.html)	Protana
PeptideSearch ^a (www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html)	EMBL
PeptIdent ^{a,b} (www.expasy.ch/tools/peptident.html)	ExPASy
ProFound ^a (prowl.rockefeller.edu/PROWL/prowl.html), (also at service.proteometrics.com/prowl/profound.html)	Rockefeller University and Proteometrics
Peptide Sequence or Peptide Sequence Tag Query	
Mascot ^a (www.matrixscience.com)	Matrix Science Ltd.
MS-Seq ^a (prospector.ucsf.edu)	UCSF
PepSea ^a (195.41.108.38/PepSeaIntro.html)	Protana
PeptideSearch ^a (www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html)	EMBL
TagIdent ^a (www.expasy.ch/tools/tagident.html)	ExPASy
MS/MS Ion Search Analysis	
Mascot ^a (www.matrixscience.com)	Matrix Science Ltd.
MS-Tag ^a (prospector.ucsf.edu)	UCSF
PepFrag ^a (prowl.rockefeller.edu/PROWL/prowl.html)	Rockefeller University
Sequest ^c (fields.scripps.edu/sequest)	Scripps Research Institute
Sonars MS/M ³ (www.proteometrics.com)	Proteometrics
^a Available for data search over the Internet.	
^b PeptIdent2 (now called SmartIdent) is a completely different tool from PeptIdent. PeptIdent2 (27) utilizes a rather unique approach involving a generic algorithm using a training set of protein mass spectra. However, it is not yet available on the publicly available ExPASy server.	
^c Sequest is commercially available through ThermoFinnigan, San Jose, CA, USA.	

(47,48), has been effectively used for this purpose (56). CIDentify can deal with the peculiarities and ambiguities of the de novo sequence interpretation process, such as presence of unsequenced dipeptide masses, mass identity of certain dipeptides with some amino acids (e.g., Asn equals the sum of two Gly residues), etc.

2. *The presence of unexpected co-translational or post-translational modifications as well as chemical modifications that are artifacts of sample handling.* The search engines are, in general, capable of handling expected modified amino acid residues. The exact way the modified amino acids are handled varies with search engines. The Mascot search engine, for example, allows searches with two types of modifications: fixed and variable. For fixed modifications it uses the

appropriate mass for the modified residue(s) for every occurrence of the specified residue(s). Usually any chemical modifications introduced as part of the sample preparation process (e.g., modification of cysteine residues) can be considered as a fixed modification. Mascot considers variable modifications as those that may or may not be present and tests all possible arrangements of variable modifications to find the best match. For example, if oxidized methionine is chosen as a variable modification for a peptide containing 2 methionine residues, it will test for a match with the experimental data for that peptide containing 0, 1, or 2 oxidized methionine residues. This is particularly useful for modifications like phosphorylation. However, the choice of a number of variable modifications prolongs the search

time considerably. Also, allowing too many possible modifications in database search programs or the *in silico* testing of all possible modifications tends to generate an increased number of random (false-positive) matches. The major difficulty with co-translational and post-translational modifications arises from the fact that the information that a residue is an actual or potential site for modification is not enough; it should be possible to equate the modification into a specific mass difference value. A software tool called FindMod (www.expasy.ch/tools/findmod), which can be applied for high-throughput determination of protein post-translational modification from peptide mass fingerprint data, has been developed (58). It examines the data for mass differences between empirical and theoretical peptides to determine the post-translational modification and can predict the amino acid residue in the peptide that might carry the modification. A total of 29 different rules were developed for this from 5153 incidences of post-translational modifications documented in the SWISS-PROT database covering 22 types of modifications.

3. *The peptide is produced by an unexpected or nonspecific cleavage.* For sequence query or a MS/MS ion search using tandem mass spectrometric data, in such cases, the “no enzyme” option for enzyme cleavage can be used that will involve search of each protein sequence in the database for peptides which meet other criteria of the search. However, this will increase the number of peptides tested by several orders of magnitude with a corresponding increment in the number of random matches. For peptide mass fingerprint analysis, the requirement for the specificity of an enzyme is absolute. The high degree of specificity of a MS/MS ion search signifies that the choice of an enzyme is not essential.
4. *The quality of the spectrum is poor or the spectrum originates from a non-peptide contaminant.* Sometimes, the quality of the spectrum can be improved by purification and concentration of peptide samples prior to mass spectrometric analysis.

Searching Different Types of Sequence Databases

To achieve maximum throughput, it is not practical to search every set of mass spectrometric data against the largest sequence database incorporating the widest set of modifications and with no enzyme specificity. The goal should be to identify proteins from as many spectra as possible using simple searches against a small but high-quality database like SWISS-PROT. It should be noted that many entries in SWISS-PROT are consensus sequences. The variant sequences are part of the annotations, and as a result, are missed by the search engines. Spectra that cannot be identified can then be searched against increasingly larger databases (e.g., non-redundant “nr” protein sequence database compiled by NCBI: the National Center for Biotechnology Information, Bethesda, MD, USA; the non-identical MSDB protein sequence database compiled by the Proteomics Department, Imperial College, London, UK; the expressed sequence tag database dbEST compiled by NCBI, etc.) with gradually more exhaustive search parameters. A good collection of information on relevant sequence databases and their downloading can be found in Reference 11 and at www.matrixscience.com/help/database_help.html.

Searching species-specific protein sequences that represent open reading frames derived from genomic sequences is generally very fast. For searching nucleotide databases, many of the search engines are capable of translating the sequence into all six reading frames on the fly. The problem with the dbEST database is the fact that a large number of peptide matches are found because of the redundant nature of the database. Also, the short sequences in this database do not often lead to the grouping of peptide matches into protein matches. To address this problem, the UniGene index (www.ncbi.nlm.nih.gov/UniGene) from NCBI has been used (12) to simplify the search results. The UniGene index has been created by BLAST analysis of GenBank sequences against themselves to cluster them into gene families. The UniGene data file for several species currently available can be downloaded from NCBI. Peptide sequence tags as well as MS/MS ion searches have also been performed on continuous raw genomic sequence data like that of the human genome (12,38). It is expected that some peptides cannot be matched because of the current “draft” nature of the sequence. Another limitation of such an approach is that peptide sequences spanning the boundaries of adjacent exons will be missed. This can be circumvented if multiple peptides are identified from a protein. However, because of the length and complexity, the standard tabular report formats used by search engines are not suitable for handling the search results. In one approach (12), the output from the peptide match results are obtained as an EMBL/GenBank format feature table. This report may not look very user-friendly, but the advantage is that it can now be read into a standard genome browser to view it. These investigators have used Artemis (www.sanger.ac.uk/Software/Artemis), a JAVA-based genome browser developed and distributed by the Sanger Centre, Cambridge, UK (Figure 1). An in-depth review on matching peptide mass spectra to EST and genomic databases for protein identification has been published recently (11).

Further Development of Software Tools for Protein Identification Stage

The software and informatic tools described above for identification of proteins from mass spectrometric data cover experimental approaches that are in common use at the present time. However, conceptual and technological advances are being continuously made on the experimental front and these are driving the development of novel software tools. A relevant example is quantitative analysis of complex protein mixtures by differential isotope labeling of peptides followed by LC-MS/MS analysis, which is currently emerging as a powerful tool in proteomics. This technology has the potential for high sensitivity, high coverage, and high throughput. Isotope-coded affinity tag (ICAT) reagents (28) developed for this purpose are now commercially available. Sometimes software tools in current use can be minimally modified to accommodate newer techniques like the ICAT technology. For example, the Mascot search engine now allows analysis of ICAT experiments. When the ICAT box is selected, searches are limited to cysteine-containing peptides, and it also allows addition of heavy and light ICAT tags to the list of variable modifications. Quite often the handling of such modifications at the level of the software may not be as simple as it might seem. It has been suggested that the same peptide con-

original sample now needs to be reconstructed by combining and analyzing potential candidates from multiple MS or MS/MS runs. In essence, the data analysis needs to reverse the fractionation steps and relate the data to the original sample.

Despite the availability of powerful mass spectrometric data search engines for identification of proteins in an analyte, there is an enormous need for supporting software tools for storage and management of the results as per the requirements described above. A specific example relates to tandem mass spectrometric methods for the direct identification of the components in a complex mixture of proteins, the so-called shotgun proteomics approach. In this application, the sample is digested directly with trypsin followed by fractionation of the peptides using single or multidimensional LC. Eluting peptides are analyzed by MS/MS and the Sequest search engine is used to identify the proteins present in the sample by correlating the MS/MS spectra with sequences in a target database. Since Sequest functions at the level of individual spectra, additional tools are necessary to assemble information at the level of proteins identified and to improve the speed and precision of data analysis. Sequest uses a flat file, directory-based storage mechanism for its output, with relevant information dispersed through various files. Also, the Sequest Summary tool can only coordinate the results from a single LC-MS/MS experiment. More functional summary and comparison tools such as DTASelect and Contrast have been recently developed (55). The former assembles Sequest identifications and highlights the most significant matches, while the latter compares DTASelect results from multiple experiments. Further information can be obtained through the Web site DTASelect and Contrast: Informatics for Shotgun Proteomics (fields.scripps.edu/DTASelect). We (D.N.C. and I.M.) have also faced similar challenges trying to collectively summarize Sequest results from large numbers of MS/MS runs. To overcome these limitations, we developed parsing and loading tools to move the flat file search results (currently consisting of almost 10 million parsed flat files) to a relational Oracle database. We then developed a custom application to query the entire dataset in a variety of ways. This application supports the quick and easy reconstruction of results derived from multiple LC-MS/MS runs. In addition, it contains comparison tools that allow scientists to quickly identify unique or common proteins (or peptides), across multiple runs and from searches against different protein databases.

As the need for such robust data management and analysis tools are becoming increasingly obvious, mass spectrometry instrument manufacturers and database search engine providers have started to develop relational database management tools to support their products. For example, a rapid, automated, data archiving and retrieval system (RADARS) for high-throughput proteomics mass spectral data processing and storage has been recently described (21). The system can identify proteins by database correlation of unprocessed mass spectrometry data files followed by storage of the processed data in a relational database for customized reporting.

DOWNSTREAM PROTEIN INFORMATICS TOOLS

The downstream protein informatics tools make the proteome profiling data comprehensible. This involves further

analysis of the data involving data mining and predictions based on computational biology. Many of these are beyond the scope of this article. Possibly the best collection of tools for analysis of proteomes at all levels is the publicly available ExPASy (Expert Protein Analysis System) proteomics server (www.expasy.ch) of the Swiss Institute of Bioinformatics. It is dedicated to the analysis of protein sequences and structures as well as 2-D gels.

Integration of Proteome Profiling Data

Large-scale and high-throughput identification of proteins in biological systems is now a reality. Once the proteome profile for a particular tissue or organism is known, informatic tools are necessary to understand the functional role of many of these proteins in the biology of the organisms they originate from. It is also necessary to link to the already existing biological knowledge base. A key requirement is the availability of comprehensive, logical, and user-friendly databases that allow the large amounts of data to be used efficiently. One approach for proteomes identified from 2-D gels involves the integration of 2-D gel databases (59). Several laboratories have built master gels of proteins identified on 2-D gels and a corresponding database containing specific data about the identified proteins. Several such databases are available on the World Wide Web. The WORLD-2DPAGE Index to 2-D PAGE databases and services (www.expasy.ch/ch2d/2d-index.html) contains extensive references to 2-D database servers, as well as to 2-D gel related servers and services. Details about creating a 2-D database on a Web server by generating the necessary programs and index files have been described (3,33). The Make2ddb package used for this purpose, which allows one to create a 2-D database with search functions, such as by accession number, by description, by author, and by clicking on a spot, similar to SWISS-2DPAGE, can be obtained freely (www.expasy.ch/ch2d/make2ddb.html). A computer-assisted visual method, Flicker, for comparing 2-D gel images across Internet databases has also been described (39). Mining of protein data from 2-D gels and the usefulness of a laboratory information processing system (LIPS) have been reviewed (45).

Careful evaluation of experimental as well as predictive and computational proteomics data and their integration with the currently available biological knowledge is the foundation for selection of drug (Figure 2), diagnostic, and vaccine targets, understanding of genetic diseases at the molecular level, and elucidation of biological pathways and networks. The key element for characterization of the ever-increasing number of predicted or confirmed protein sequences is annotation or the assignment of function to these proteins. Large publicly available databases are necessary to record the massive amounts of data that will be collected from proteomics experiments carried out with different organisms by different research groups. A related issue is development of abilities to create such massive databases automatically from experimental data. Integrated databases that contain annotated genes and proteins with all relevant information collected from the scientific literature are of paramount importance to gain substantially from the massive amounts of proteome data currently being generated. Examples of such well-annotated data are the NiceProt View of the entries in the SWISS-PROT database (www.expasy.ch) as well as the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* proteome databases (www.proteome).

com). The Protein Information Resource (PIR), in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database, also produces a comprehensive and expertly annotated protein sequence database, the PIR-International Protein Sequence Database. This database and search tools are available on the PIR Web site (pir.georgetown.edu) and the MIPS Web site (www.mips.biochem.mpg.de). Another useful resource on the human proteome is the International Protein Index (IPI), available from the European Bioinformatics Institute (www.ebi.ac.uk/IPI/IPIhelp.html). IPI is assembled by automatic mapping of human protein sequence information taken from the following five data sources: SWISS-PROT, TrEMBL, RefSeqNPs, RefSeqXPs, and Ensembl. TrEMBL is an automatically annotated supplement to SWISS-PROT. RefSeqNPs and RefSeqXPs are a reference set of protein sequences and a reference set of *Homo sapiens* model proteins provided by the genome annotation process maintained by NCBI, respectively. Ensembl is a joint project between EBI and the Sanger Centre to develop a software system that produces and maintains automatic annotation on eukaryotic genomes. The aim of the IPI is to provide a minimally redundant yet maximally complete set of human proteins. Each IPI entry consists of a cluster of related entries from the constituent databases along with a sequence (FASTA format) and a description taken from the master entry. With advances in functional proteomics, the

availability of databases focused on analysis of networks and fluxes in living systems will also be an important issue. Proteomics is one of the major providers of information for systems biology approaches (34,35), which are focused on generating an integrated understanding of biological processes at the molecular level using mathematical tools to extract information. Such an understanding is crucial for the discovery of novel markers for diseases, drugs, and vaccines that together will aid in the early detection as well as effective prophylactic or therapeutic treatment.

OTHER RELATED PROTEOME INFORMATICS TOOLS

LIMS in Proteome Analysis

LIMS software packages have been used extensively in the analytical laboratories to manage workflow, samples, and results in a systematic and organized way. Integration with analytical instruments, laboratory robotic devices, and a result reporting structure has been some of the central features of LIMS packages. LIMS packages have excelled in handling large-scale sample processing and analysis needs in many industrial and pharmaceutical laboratories performing high-throughput compound screening. However, until recently, their utility in the proteomics laboratory has not been fully realized. Advances in high-

Figure 2. Central role of proteome profiling in our understanding of biological models and drug discovery. Biological models are developed based on our current knowledge in biology organized in genomic and proteomic databases [6,7]. Biological models are experimentally tested using proteome profiling [1] as well as other technologies such as gene expression profiling, functional assays, and genetic technologies. Proteome profiling technologies rely heavily on knowledge from currently available proteomic and genomic databases [2,4] as explained in this article. In return, data from proteome profiling (protein identification, quantitation, post translational modification, interactions, etc.) greatly enrich the existing knowledge in these databases [3,5]. Structure, function, expression, and interaction data from proteome profiling form the basis for the discovery of drug targets [8,9,10] whose properties can be modulated in a pharmacologically beneficial manner by drugs identified in the vast compound libraries of the pharmaceutical industry.

throughput protein production and fractionation technologies, as well as high-throughput, large-scale MS and MS/MS for protein and peptide characterization have brought the benefits of flexible and customizable LIMS packages for the proteomics laboratory into consideration. LIMS packages are expensive to acquire, customize, and maintain both in terms of financial and human resources. Thus, they usually remain in the realm of large pharmaceutical or academic laboratories with ample information technology and bioinformatics resources.

In our experience (D.N.C. and I.M.) with developing a proteomics LIMS (a collaboration with Applied Biosystems, Foster City, CA, USA), we have a better understanding of the strengths and limitations of the technology. However, we are convinced that the time and effort we have spent to customize the basic SQL*LIMS package (Applied Biosystems) to reflect the proteomics workflow and address the requirements of proteomics bench scientists was worthwhile. The LIMS package not only maintains the traditional functionality of similar systems, but with the appropriate extensions and integration into a Web-JAVA (J2EE) based architecture also allows for the collection, management, and analysis of results from a variety of custom-built proteomics and bioinformatics applications. The LIMS relational database provides full support for storing, reporting, and analyzing proteomic information, data, and results using both a traditional client-server architecture as well as a thin Web-based client. WorksBase software from BioRad Laboratories (Hercules, CA, USA) is another powerful LIMS package for proteomics.

CONCLUSION

In a proteomics project, large amounts of data are generated very rapidly. The information generated by all of the processes described for proteome profiling needs to be collated and linked to existing information in scientific literature. The availability of proper data analysis, visualization aids, mining, and summary tools as well as integration of results with 2-D gel and LIMS packages is very important. One of the most important issues facing proteomics research is that of cross-platform operability of software tools, data integration, and data sharing. A modern proteomics laboratory uses data from different sources, generated using different instruments and different scientific software. Several groups such as the Global e-MS Consortium (information available through link in Micromass UK Limited's home page at www.micromass.co.uk) and the Interoperable Informatics Infrastructure Consortium (I3C; www.i3c.org) are actively working on these issues related to development of software tools capable of operating across different platforms as well as common protocols and technologies for data exchange and knowledge management.

Another informatics challenge stems from the widely recognized notion that a gene product can have more than one molecular function, be used in one or more biological processes, and may be associated with multiple cellular components. Because of this, the major challenge for automated and large-scale annotation of protein function is the development of a standardized vocabulary for names of proteins, genes, and their functions. The Gene Ontology™ (GO) Consortium (www.geneontology.org) is working to develop a dynamic controlled vocabulary applicable to all organisms to manage the accumulating and

changing knowledge of genes and the roles of their corresponding proteins in cells (6).

Author's Note: All of the universal resource locators (URLs) listed in this article were tested and found to be functional in February, 2002.

ACKNOWLEDGMENTS

We would like to thank Dr. Robert Zagursky and Dr. Subramonia Pillai of Wyeth Lederle Vaccines for their encouragement and support. We would also like to thank Terri Ward and Eileen Shirley of the Wyeth Lederle Vaccines library for their help in obtaining several literature references cited in this article. Thanks are also due to our colleagues at the Genetics Institute—Wyeth Ayerst Research (WAR), Wyeth Lederle Vaccines, and CNS Proteomics groups for providing requirements and advice during software development, and the members of the Genomics/Bioinformatics group at WAR, Gary VanDomselaar, Patrick Cody, Dr. Steve Howes and Dr. Charles Richard. We would also like to acknowledge the help of Dr. John Cottrell of Matrix Science Limited for providing Figure 1.

REFERENCES

1. Abbott, A. 2001. Workshop prepares ground for human proteome project. *Nature* 413:763.
2. Anderson, N.L., J. Taylor, A.E. Scandora, B.P. Coulter, and N.G. Anderson. 1981. The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.* 27:1807-1820.
3. Appel, R.D., C. Hoogland, A. Bairoch, and D.F. Hochstrasser. 1999. Constructing a 2-D database for the World Wide Web. *Methods Mol. Biol.* 112:411-416.
4. Appel, R.D., D.F. Hochstrasser, M. Funk, J.R. Vargas, C. Pellegrini, A.F. Muller, and J.R. Scherrer. 1991. The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* 12:722-735.
5. Appel, R., D. Hochstrasser, C. Roch, M. Funk, A.F. Muller, and C. Pellegrini. 1988. Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* 9:136-142.
6. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25-29.
7. Beavis, R.C. and D. Fenyo. 2000. Database searching with mass-spectrometric information. In W. Blackstock and M. Mann (Eds.), *Proteomics: A Trends Guide*, p. 22-27. Elsevier, Amsterdam and New York.
8. Berndt, P., U. Hobohm, and H. Langen. 1999. Reliable automatic protein identification from matrix-assisted laser desorption/ionisation mass spectrometric peptide fingerprints. *Electrophoresis* 20:3521-3526.
9. Cahill, D.J. 2001. Protein and antibody arrays and their medical applications. *J. Immunol. Methods* 250:81-91.
10. Chakravarti, D.N., M.J. Fiske, L.D. Fletcher, and R.J. Zagursky. 2000. Mining genomes and mapping proteomes: identification and characterization of protein subunit vaccines. *Dev. Biol.* 103:81-90.
11. Choudhary, J.S., W.P. Blackstock, D.M. Creasy, and J.S. Cottrell. 2001. Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol.* 19:S17-S22.
12. Choudhary, J.S., W.P. Blackstock, D.M. Creasy, and J.S. Cottrell. 2001. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* 1:651-667.
13. Clauser, K.R., P. Baker, and A.L. Burlingame. 1999. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71:2871-2882.
14. Conrads, T.P., G.A. Anderson, T.D. Veenstra, L. Pasa-Tolic, and R.D. Smith. 2000. Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* 72:3349-3354.

15. Cottrell, J.S. 1994. Protein identification by peptide mass fingerprinting. *Pept. Res.* 7:115-124.
16. Dancik, V., T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. 1999. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 6:327-342.
17. Eng, J.K., A.L. McCormack, and J.R. Yates, III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5:976-989.
18. Eriksson, J., B.T. Chait, and D. Fenyo. 2000. A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* 72:999-1005.
19. Fenyo, D. 2000. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* 11:391-395.
20. Fenyo, D., J. Qin, and B.T. Chait. 1998. Protein identification using mass spectrometric information. *Electrophoresis* 19:998-1005.
21. Field, H.I., D. Fenyo, and R.C. Beavis. 2002. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* 2:36-47.
22. Garrels, J.I. 1989. The QUEST system for quantitative analysis of two-dimensional gels. *J. Biol. Chem.* 264:5269-5282.
23. Garrels, J.I. 1979. Two dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* 254:7961-7977.
24. GenomeWeb News. 2001. Backed by Compaq, Bruker, and Novartis, Gene-Prot opens industrial-scale proteomics facility. 26 April (www.genomeweb.com).
25. Goodlett, D.R., A. Keller, J.D. Watts, R. Newitt, E.C. Yi, S. Purvine, J.K. Eng, P. von Haller, et al. 2001. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun. Mass Spectrom.* 15:1214-1221.
26. Goodlett, D.R., J.E. Bruce, G.A. Anderson, B. Rist, L. Pasa-Tolic, O. Fiehn, R.D. Smith, and R. Aebersold. 2000. Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. *Anal. Chem.* 72:1112-1118.
27. Gras, R., M. Muller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, C. Hoogland, J.C. Sanchez, et al. 1999. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* 20:3535-3550.
28. Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17:994-999.
29. Haab, B.B., M.J. Dunham, and P.O. Brown. 2001. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* 2:RESEARCH0004.
30. Han, D.K., J. Eng, H. Zhou, and R. Aebersold. 2001. Quantitative profiling of differentiation-induced microosomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* 19:946-951.
31. Henzel, W.J., T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley, and C. Watanabe. 1993. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 90:5011-5015.
32. Herbert, B.R., J.L. Harry, N.H. Packer, A.A. Gooley, S.K. Pedersen, and K.L. Williams. 2001. What place for polyacrylamide in proteomics? *Trends Biotechnol.* 19:S3-S9.
33. Hoogland, C., V. Baujard, J.C. Sanchez, D.F. Hochstrasser, and R.D. Appel. 1997. Make2ddb: a simple package to set up a two-dimensional electrophoresis database for the World Wide Web. *Electrophoresis* 18:2755-2758.
34. Ideker, T., T. Galitski, and L. Hood. 2001. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2:343-372.
35. Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929-934.
36. James, P., M. Quadroni, E. Carafoli, and G. Gonnet. 1993. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 195:58-64.
37. Jensen, O.N., A.V. Podtelevnikov, and M. Mann. 1997. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* 69:4741-4750.
38. Kuster, B., P. Mortensen, J.S. Andersen, and M. Mann. 2001. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 1:641-650.
39. Lemkin, P.F. 1997. Comparing two-dimensional electrophoretic gel images across the Internet. *Electrophoresis* 18:461-470.
40. Lemkin P.F. and L.E. Liplin. 1981. GELLAB: a computer system for 2D gel electrophoresis analysis. I. Segmentation of spots and system preliminaries. *Comput. Biomed. Res.* 14:272-297.
41. Mann, M. and M. Wilm. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66:4390-4399.
42. Mann, M., P. Hojrup, and P. Roepstorff. 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22:338-345.
43. Moseley, M.A. 2001. Current trends in differential expression proteomics: isotopically coded tags. *Trends Biotechnol.* 19: S10-S16.
44. NCBI Education. What is Bioinformatics? (www.ncbi.nih.gov/Education).
45. Oh, J.M., S.M. Hanash, and D. Teichrow. 1999. Mining protein data from two-dimensional gels: tools for systematic post-planned analyses. *Electrophoresis* 20:766-774.
46. Pappin, D.J.C., P. Hojrup, and A.J. Bleasby. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3:327-332.
47. Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* 183:63-98.
48. Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.
49. Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551-3567.
50. Persidis A. 1999. Bioinformatics. *Nat. Biotechnol.* 17:828-830.
51. Pun, T., D.F. Hochstrasser, R.D. Appel, M. Funk, V. Villars-Augsburger, and C. Pellegrini. 1988. Computerized classification of two-dimensional gel electrophoretograms by correspondence analysis and ascendant hierarchical clustering. *Appl. Theor. Electrophor.* 1:3-9.
52. Rabilloud, T. 2002. Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* 2:3-10.
53. Rowlands, D.G., A. Flook, P.I. Payne, A. van Hoff, T. Niblett, and S. McKee. 1988. GESA—a two-dimensional processing system using knowledge base techniques. *Electrophoresis* 9:820-830.
54. Sidhu, K.S., P. Sangvanich, F.L. Brancia, A.G. Sullivan, S.J. Gaskell, O. Wolkenhauer, S.G. Oliver, and S.J. Hubbard. 2001. Bioinformatic assessment of mass spectrometric chemical derivatization techniques for proteome database searching. *Proteomics* 1:1368-1377.
55. Tabb, D.L., W.H. McDonald, and J.R. Yates, III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* (In press).
56. Taylor, J.A. and R.S. Johnson. 2001. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 73:2594-2604.
57. Taylor, J.A. and R.S. Johnson. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11:1067-1075.
58. Wilkins, M.R., E. Gasteiger, A.A. Gooley, B.R. Herbert, M.P. Molloy, P.A. Binz, K. Ou, J.C. Sanchez, et al. 1999. High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* 289:645-657.
59. Wilkins, M.R., D.F. Hochstrasser, J.C. Sanchez, A. Bairoch, and R.D. Appel. 1996. Integrating two-dimensional gel databases using the Melanie II software. *Trends Biochem. Sci.* 21:496-497.
60. Yates, J.R., III. 1998. Database searching using mass spectrometry data. *Electrophoresis* 19:893-900.
61. Yates, J.R., III, S. Speicher, P.R. Griffin, and T. Hunkapiller. 1993. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 214:397-408.
62. Zhang, W. and B.T. Chait. 2000. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* 72:2482-2489.

Address all correspondence to:

Deb N. Chakravarti
 Biotechnology Discovery Research Group
 Wyeth Lederle Vaccines
 211 Bailey Road
 West Henrietta, NY 14586, USA
 e-mail: chakradn@war.wyeth.com