In: New Developments in Robotics Research
Editor: John X. Liu, pp. 217-252

ISBN 1-59454-593-6

Chapter 7

ROBOTICS, AUTOMATION, AND STATISTICAL LEARNING FOR PROTEOMICS

***Gil Alterovitz*[1], *Ehsan Afkhami*[2], *and Marco Ramoni*[3]**
[1]Health Science and Technology/Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.
[2]Electrical and Computer Engineering, Boston University, Boston, MA, USA.
[3]Harvard Medical School and Harvard Partners Center for Genetics & Genomics, Boston, MA, USA

Abstract

During the era of the Human Genome Project [1], the emphasis was on sequencing and annotating individual genes. At that time, the number of estimated human genes was thought to be 100 thousand genes. Yet, as the human genome project draws to a close [2], recent work has decreased the estimate to between 20-25 thousand not far from the number of genes in a simple worm (i.e. *C. elegans*). Thus, the complex engineering of a human must be from other areas such as the interactions of the gene's products, or proteins. Given this, the field of proteomics has quickly been drawn to center stage. While biologists seek to study proteins, methods have been rather primitive until recently. A sudden surge of engineering and other technical talent has led this field and associated research to grow dramatically in the last couple of years.

In this chapter, the topic of proteomics is introduced to an engineering/technical audience with an emphasis on the robotics and intelligent systems technologies used in this field. These include issues in protein extraction, separation, and identification. The associated analysis algorithms and statistical learning methods are also discussed. Two case studies regarding the above topics are then explored. Lastly, the future direction of the field and its challenges are delineated. Clinical applications of proteomics such as cancer diagnosis and drug discovery are expounded upon as relevant.

Keyword: Proteomics, Robotics, Automation, Statistical Learning, Mass spectrometry

I The Promise of Proteomics

Proteins are essentially the small machines that allow an organism to function. “Proteomics,” a term introduced in the early 1990s, is a field concerned with determining the structure, expression, localization, interactions and cellular roles of all proteins within a particular organism or subset of one [3]. Yet, until recently, it was only possible to explore proteins and their function one at a time. Indeed, the key to proteomics is its intrinsic focus on parallelization and computational techniques to study myriad proteins at the same time.

Proteomics is set to have a profound impact on clinical diagnosis and drug discovery. In fact, most drugs target and inhibit the functions of specific proteins.

Proteomics has come a long way since the mid-1990s when protein networks were largely studied using 2-D gel electrophoresis (discussed later) [3]. Clinical proteomics is concerned with identifying protein networks and the intracellular interactions between proteins as applied to clinical aims [4]. The functioning of the human cell can be likened to the operation of a factory, as proteins are machines that process/deliver products and messages to other proteins via biochemical interactions. These messaging pathways or routes are essential for cellular function. As such, their malfunction can also be the cause or consequence of a disease process [4]. It is this notion that stimulated the application of proteomic technologies to: oncology [5], neurology [6], toxicology [7], immunology [8], and many other areas [9-11]. Later in the chapter, mass spectrometry methods and their proteomics applications will be outlined. With robust and high throughput features, these tools have enabled the resolution of thousands of proteins and peptide species in bodily fluids ranging from blood [12] to urine [13, 14]. Such technologies have advanced research in early cancer diagnosis as well as in Human Immunodeficiency Virus (HIV) inhibiting drugs [4, 15].

Proteomics can and does leverage some of the engineering and statistical methodology developed for functional genomics approaches [16]. However, challenges have arisen in this new field and customized solutions such as fabrication of chips for parallelization of experiments [17-24], robotics [25-31], and machine learning techniques for intelligent decision analysis [32-34] need to be engineered. Other challenges are completely new and proteome specific. For example, post-transitional modifications of proteins can be vital for cell function. In such cases, one to one correspondence does not exist between each protein and its encoding gene. This is significantly different from the relatively static nature of DNA. Since the post-translational modifications occur after the protein is created (based on the genetic blueprint), such modifications cannot be seen via traditional genomics approaches.

The Human Genome Project has demonstrated that speed, cost, and precision are the underlying factors in any large scale biological endeavor and that technological hurdles can be overcome with novel engineering approaches. Higher throughput and sensitivity are requirements of technologies aiming to capture quality snapshots of cellular activity. It is with this aim that academia and industry are pushing ahead in the automating processes such as robotic sample preparation [35], alternative readouts for protein interactions [36-38], and microfluidics [39]. Current instrumentation is far from optimal, however, partly because manufacturers have not yet had the necessary lead time to build systems perfectly tailored to protein analysis [40].

In addition to sensitivity and throughput considerations, there are many data analysis challenges inherent in representation and interpretation of experimental results. Methods aimed at meeting these problems are largely grouped under bioinformatics, a multidisciplinary discipline, absorbing methods in computer science, signal processing, statistical inference, and other engineering-related fields. Algorithms such as the Basic Local Alignment Search Tool (BLAST) [41] have been developed for automated protein identification. Yet, more intelligent decision making algorithms are needed to improve detection of post-transitional modifications in MS spectra, peptide mass fingerprinting, and electrophoresis image analysis.

II Introduction to Molecular and Cellular Biology

This section summarizes some of the core molecular and cellular biology concepts that underlie the study of proteomics [42, 43]. At the microscopic level all living species are composed of cells, each with differing complexities and functionalities. However, all cells, whether they are human or bacteria, share some common functional parts. Cells typically consist of an outer wrapper called the cell membrane, with a watery fluid inside called the cytoplasm. Cytoplasm is approximately 70 percent water; the other 30 percent is filled with proteins made by the cell, along with smaller molecules. At the center of the cell is the nucleus, a compartment that holds the master controller of the cell: deoxyribonucleic acid (DNA). DNA serves as the blueprint for proteins by encoding the genes that are used to create them.

A DNA: A Blueprint for Proteins

DNA guides the cell in its production of new proteins. The DNA in a cell can be thought of as encoded message. Rather than a binary (0 or 1) or alphanumeric (26 possibilities per letter) message, each ‘letter’ in the message is one of four different nucleotides, or bases. The four bases in DNA's alphabet are: adenine (A), cytosine (C), guanine (G) and thymine (T). Each message has built-in error detecting code via a redundant message that is encoded in parallel. That is, each base has a complimentary base, so as to form a pair bonded together at each ‘letter’ position. Adenine and thymine always bond together as a pair, and cytosine and guanine bond together as a pair. The pairs link together like rungs in a ladder as shown in Figure 1. In this example, the top message is ACGTACC from left to right. Assuming no errors, the bottom ‘redundant’ sequence simply follows the aforementioned nucleotide pair bonding rules.

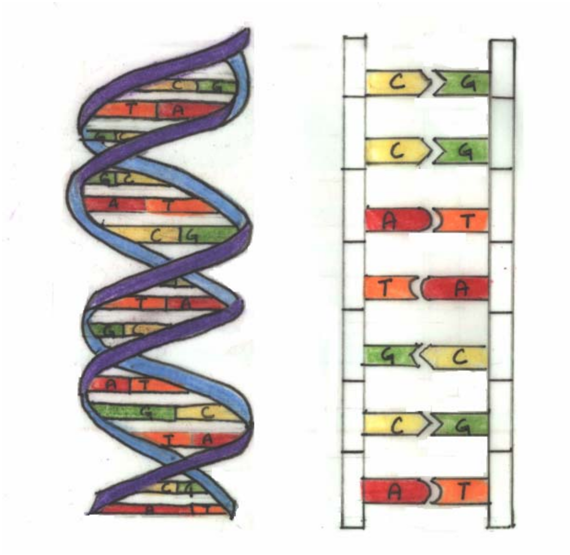


Figure 1. The DNA

A gene is a section of DNA that acts as a template for the ultimate structure of a protein (see Figure 2). Consequently, the DNA can be thought of as a consisting of a sequence of genes that encode for all that an organism is made of. DNA in the human genome is arranged into 24 distinct chromosomes, physically separate molecules that range in length from about 50 million to 250 million DNA base pairs.

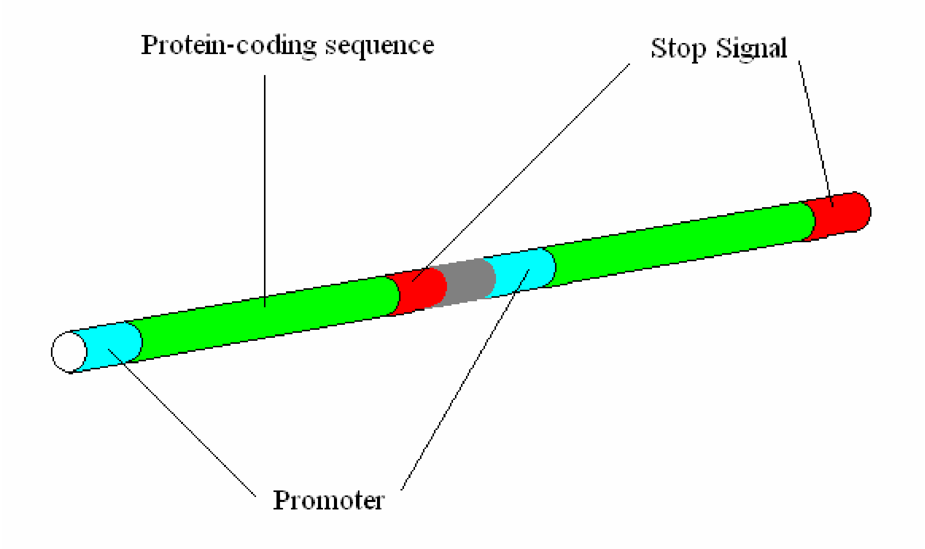


Figure 2. A Gene

The human genome project has established all 3 billion of the base pairs in a typical human's DNA [44]. These base pairs are estimated to encode for 20,000 - 25,000 protein-coding genes [2]. Genes, however, comprise only about 2% of the human genome [45]. The remainder consists of non-coding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

B Proteins: Molecular Machines

A protein is any chain of amino acids. Amino acids are organic compounds containing an amino group (NH2) and a carboxyl group (COOH). Proteins are made of amino acids by stringing together as many as few hundred amino acids in a very specific and unique order.

Figure 3 illustrates the chemical structure of two amino acids. It can be seen that the top portion of each one is the same, a common feature of all amino acids. The functional chain at the bottom (the circled H and CH3 in these two amino acids) is the variable region from one amino acid to another.

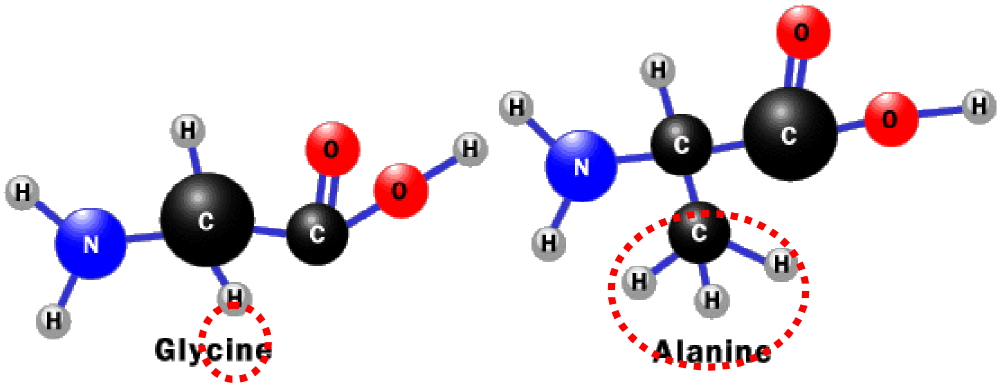


Figure 3. Amino Acids

The covalent linkage between two adjacent amino acids in a protein chain is called a peptide bond. For this reason, proteins can also be referred to as polypeptides. Regardless of the specific amino acids from which it’s made, the polypeptide has an amino (NH2) at one end and carboxyl group (COOH) at its other end.

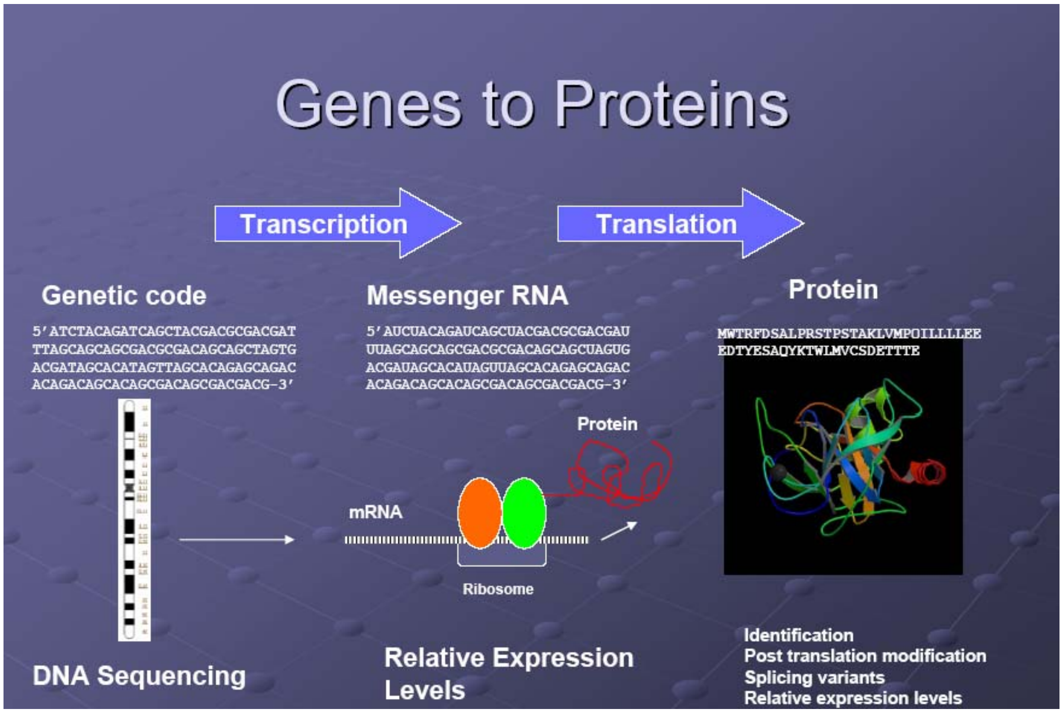
The chain of amino acids folds into a unique shape. The chemical makeup and shape of a protein allow it to participate in chemical reactions necessary for the cell’s sustenance. For example, the sugar maltose is made from two glucose molecules bonded together. The maltase protein is shaped in such a way that it can break the bond and free the two glucose molecules.

Proteins can be broadly organized and categorized by their functions and structures, performing structural roles, handling metabolic chores, and participating in signaling pathways, to name a few. They can be thought of as being made of modular units (motifs/domains) that confer specific properties and functions. These are recognizable amino acid sequences that show similar properties or functions when they occur in a variety of proteins. The significance of motifs and domains for proteomics is that they represent the transformation of peptide sequence to protein functions. The cellular roles of proteins with unknown functions can be predicted by the occurrence of recognizable motifs within them. Said another way analytical proteomics can define sequence and sequence can define biological function [46]. It should be noted that approximately 40% of the human genome encodes proteins with no known function [47]. Assigning functions to these proteins and their interactions is one of the challenge of proteomics [48].

C The Central Dogma: From DNA to RNA to Protein

There are several steps that need to occur for the information encoded in DNA, made up of only four unique nucleotides, to be transformed into a protein, containing 20 unique amino acids. This process is part of what is known as the central dogma of molecular biology [49].

To create a protein (see Figure 4), the cell must first transcribe the gene in the DNA into messenger ribonucleic acid (mRNA). If DNA is like the permanent paper blueprint, then RNA is like the current day's memo on how to implement it. In fact, DNA is inherently a more stable molecule than RNA. Thus, DNA is suitable for long-term storage while RNA is used for communicating quick messages regarding the cell's current protein needs. The transcription is performed by a protein called RNA polymerase. RNA polymerase binds to the DNA strand at the so-called promoter site, unlinks the two strands of DNA and then makes a complementary copy of one of the DNA strands into a RNA strand. One additional complexity that can occur here is alternative splicing [42] in which variable segments of a gene's message can be spliced together to form the ultimate messenger RNA.



Genes to Proteins

Transcription

Translation

Genetic code

5′ATCTACAGATCAGCTACGACGCGACGAT
TTAGCAGCAGCGACGCGACAGCAGCTAGTG
ACGATAGCACATAGTTAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′

Messenger RNA

5′AUCUACAGAUCAGCUACGACGCGACGAU
UUAGCAGCAGCGACGCGACAGCAGCUAGUG
ACGAUAGCACAUAGUUAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′

Protein

MWTRFDSALPRSTPSTAKLVMPOILLLLEE
EDTYESAQYKTWLMVCSDETTTE

DNA Sequencing

Relative Expression Levels

Identification
Post translation modification
Splicing variants
Relative expression levels

Figure 4. The Central Dogma: From Genes to Proteins

In any case, the strand of messenger RNA then goes from the nucleus to a ribosome, a compartment within the cell that translates the RNA sequence to a protein. To generate the right amino acids, a ribosome takes the nucleotides in sets (referred to as a codons) of three to encode for each of the 20 amino acids. For every three base pairs in the original DNA sequence, one amino acid is generated for creation of a protein. Because DNA consists of four different bases, and because there are three bases in a codon, there are $4^3 = 64$ possible combinations for a codon. Yet, since there are only 20 possible amino acids, some redundancy exists in the codon codebook and several different codons can encode for the same amino acid.

The result is that the codons are translated into a string of the amino acids that are strung together to form a long chain. When the last codon, the stop codon, is reached, the ribosome releases the chain to form the final protein. It quickly folds into its characteristic shape, floats free, and begins performing its function in the cell.

Once translated, posttranslational modifications can alter protein. These changes are the finishing touches put on the protein to be able to perform its function and may occur at any time during the protein's life cycle. One such posttranslational modification critical to cell regulation is phosphorylation, a process by which a phosphate group is added to the protein.

Proteomics technologies are used to identify posttranslationally modified proteins, something that cannot be analyzed using genomics since these changes occur after the translation process. For example, one proteomics study on phosphorylation [50] has examined the effects of phosphotyrosine-containing proteins in lung cancer tissues. The phosphorylation of proteins in the lung tissue showed a high level of correlation with post-surgery prognosis. They concluded that tyrosine phosphorylation of the proteins involved in regulating cell adhesion were correlated with the survival. This finding suggested that this phosphorylation may perturb cell-cell adhesion and activate tumor invasion.

D From Genome to Proteome

At the DNA level, each cell contains all the information necessary to make a complete human being. However not all genes are expressed in each cell. Genes that encode for proteins essential to basic cellular functions are expressed in virtually all cells, whereas those with highly specialized functions are expressed only in specific cell types. Every organism has one genome but many proteomes, thus the proteome in any cell represents some subset of all possible gene products.

The recent completion of the human genome sequence has provided evidence that the human genome encodes between 20,000 and 25,000 genes as noted earlier. Interestingly, this is only about slightly larger than the approximately 19,000 genes contained in the worm (Caenorhabditis elegans) genome [51]. In view of the tremendous differences in the complexity of the human organism compared to the worm, the value of proteomic over genomic approaches becomes evident. That is, the complexity of the human organism must lie in the diversity of human proteins and their interactions rather than in the static human genome.

Genomics focuses on the statistic structure of the DNA and aims to determine the DNA sequence of various organisms and differentiating between individual's sequences. The next level of complexity is the area of functional genomics which deals with the amount of mRNA transcription in cells. Cells use alternative splicing to produce different transcripts from the same gene; this means that there isn't a one to one relationship between the genome and the transcript. Although mRNA profiling through microarrays offers immense potential for the understanding of molecular changes that occur during biological processes including disease progression, it does not capture mechanisms of regulation involving changes in cellular localization, sequestration by interaction partners, proteolysis and recycling. Studies in yeast have shown that there is a weak correlation between mRNA levels and protein expression. In fact, mRNA levels in some genes were the same value as others while the protein levels varied by more than 20-fold [52]. The level of any protein in a cell at any given time is controlled by a number of variables:

- The rate of transcription of the gene
- The efficiency of translation of mRNA into protein
- The rate of degradation of the protein in the cell

Proteomics is the next layer of analysis.

Any protein, though a product of a single gene, may exist in multiple forms at any given time. Most proteins exist in several modified forms which affect protein structure and function. The status of the proteome within a cell reflects all the cell's functions. The challenge of proteomics is detecting many relatively low abundant proteins that play a role beyond general cell upkeep and which may exist in multiple modified forms. In recent years, proteins with specific amino acid sequences, structures, functions, concentrations, and post-transitional modifications have all been explored [53].

Proteomics encompasses four major applications. Mining is the process of identifying and cataloging as many proteins as possible directly rather than inferring them from gene expression. Protein expression profiling is the identification of protein abundance while the organism is in a specific state. This could be exposure to drug or a disease state. Protein-protein network mapping is concerned with how proteins interact with each other within a cell. These interactions can be permanent or transient. Lastly protein modification studies strive to identify how and where proteins are modified.

Even minute changes to proteins can cause major changes in function with pathological consequences. For example, a change in just one amino acid in one type of polypeptide chain can result in sickle cell anemia, a devastating hemolytic disease that often results in death as a result of abnormal red blood cell function and recurrent clotting episodes [54].

III Technologies & Automation in Proteomics

The move towards robotics and automation in the life sciences has been underway for nearly 20 years [55]. The growth of this research area is illustrated in Figure 5 below. Using the Medical Subject Heading (MeSH) database and the PubMed citation database [56-58], the number of annual research articles in were calculated within several topics as a proxy for research activity. These included: automation, robotics, and biomedical engineering-related fields. These were compared to all research articles that appeared in the index annually. For each subcategory, the y-axis is normalized to the number of articles published in 2003 within that subcategory (100%). Thus, the growth of the various fields can be compared to the overall growth of research papers during the decade 1993-2003. In particular, all of the technologies related to automation, robotics, and biomedical engineering-related fields grew at a similarly spectacular rate of approximately 3-5 fold, while the overall citation index only grew by around 1/3. The graph shows that this growth gives no sign of saturation.

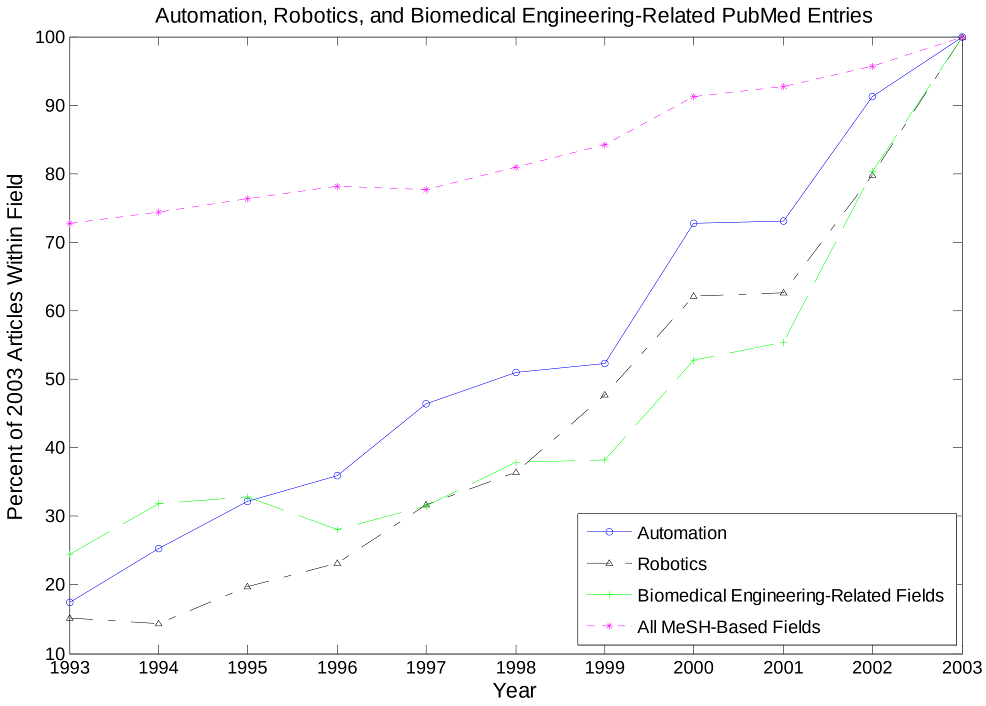


Figure 5: Automation, Robotics, and Biomedical Engineering-Related Papers are growing at a much faster rate than the papers in all fields in the PubMed database.

Researchers are looking to robotics to search entire proteomes for potential targets for treatment. Robotics can increase throughput, eliminate sample contamination, reduce human error, and perform repetitive processing. In particular, the high-throughput demands of the pharmaceutical industry for drug screening have resulted in an increased need for automotive approaches to supplant historically manual techniques.

Automation has become common place in all stages of an assay from sample preparation (see Figure 6) to processing, analysis, and information management (see Figure 6). Bench-top automated liquid handling and sample dispensing systems are becoming widely available. Miniaturized pipetting robots, though expensive, save researchers money simply by using less (20 nanoliters) of the costly reagents used in biomedical research. Automated protein purification is now possible with microfabrication technology developed for semiconductor research in the form of “chips” with microscopic channels [55]. Small electric currents or vacuum-based pressure techniques can used to conduct the flow of fluids. Electrophoresis gel imaging, robotic gel cutting, and mass spectrometry sample plate loading are other examples of automation [59-61].

To extract useful information from terabytes of data gained during the automated process, information management systems specific to the life sciences have been created. Laboratory Information Management Systems (LIMS), as they are typically called, are designed to mirror the natural work flow of the laboratory, integrating manual and automated processes. For example, robotic platforms can track a sample and its accompanying data through various processes [55]. An example of LIMS is Nautilus, a proprietary software suite where data is put into extensible markup language (XML) format, a standard in many industries for storing data structures [62].

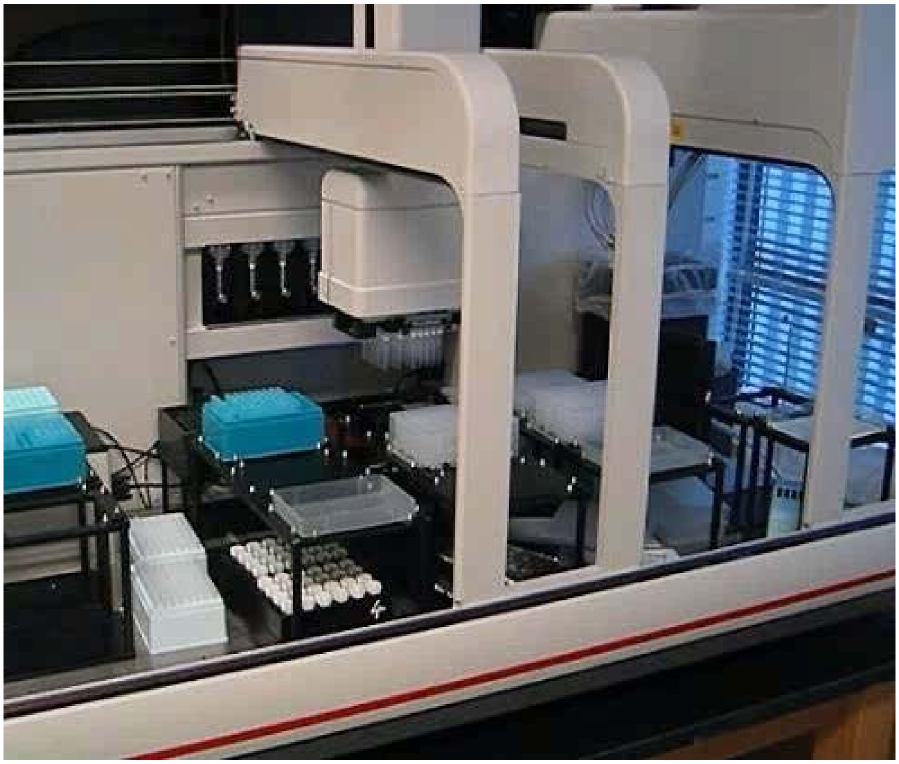


Figure 6. Sample Preparation Robot

Automation and robotics also have introduced some novel problems which have opened up new avenues for research [55]. Downtime for reconfiguration or replacements can significantly hinder throughput. Research from fault tolerant networks, redundant machinery, and/or parallelization can prove useful here [63-69]. Integration between machinery from various vendors is another issue in lab automation. A trade off exists between buying whole systems from one vendor (where individual components may not meet all specifications) versus for separate vendors (where intercomponent integration may be more difficult).

A Analytical Polypeptide Separation

Mass spectrometry has not been able to identify whole proteins solely based on their molecular masses. This is due to the fact that mass spectrometry measurement accuracy decreases as the protein mass increases, multiple proteins have similar masses, post-transitional modifications complicate the assignment based on protein mass, and lastly, not all proteins are amenable to intact mass measurements [70]. More discussion of some of the statistical issues involved is presented in the next section.

The essence of analytical protein identification centers around the following: most peptide sequences of approximately six or more amino acids are largely unique within the proteome of an organism [46]. This will result in identifying a protein based on the identification of a hexapeptide (i.e. a peptide consisting of six amino acids). The confidence in this match is increased if multiple partial pieces of the entire protein can be matched.

Thus, proteins can be identified via a multi-step process. First, they are cut into small pieces (i.e. small peptides) though a digestion process and these small pieces can then be identified via mass spectrometry (MS) to a high degree of accuracy (unlike the entire protein). A database can then identify which protein these small peptides originated from.

Yet, even before the digestion process and mass spec analysis, a number of steps are needed to facilitate analysis. Proteins must be extracted from biological samples such as a piece of tissue or cultured cells. The next step is to separate the proteins contained within the tissue. The most popular protein separation methods are 2-D gel electrophoresis (e.g. sodium dodecyl sulfate-polyacrylamide gel electrophoresis, or SDS-PAGE for short), preparative isoelectric focusing (IEF), and high performance liquid chromatography (HPLC). HPLC and MS (HPLC-MS) is a combination that has lent itself well to automation and it is thus expected that HPLC will likely dominate polypeptide separation in the long run (though 2-D SDS-PAGE is still prominent today [53]).

In 2-D SDS-PAGE, proteins are separated first by their isoelectric point followed by separation according to molecular weight. The result is the separation of proteins into spots on a gel containing sample proteins. The intensity of each spot is proportional to the protein abundance. The stained gel image can be analyzed using imaging analysis techniques and a section of the gel containing an isolated protein can be cut out for further analysis by other methods such as mass spectrometry. Two or more samples from differing cellular states (diseased and normal) can be compared to identify relevant proteins.

Integrated systems for performing the above tasks are currently being made available. These systems include: robotic sample preparation, 2-D gel electrophoresis, gel extraction via precision robots, ionization labeling, and MS peptide fragments analysis. In these systems, data generated from all the instruments are represented in a user friendly graphical user interface (GUI) [71] for easy analysis. These systems are crucial to high throughput, in some instances increasing processing power by 5 fold [72]. A shortcoming in these systems stems from the fact that samples are typically treated in a homogenous fashion with no feedback control mechanism. For example, a lab technician doing a gel protein digestion can account for the spot intensity by adjusting the amount of protease (an enzyme used to cleave the protein into peptides) and re-suspension volume based on the sample. However, intelligent systems are not yet available to make such decisions [72].

Electrophoresis's application is limited due to its small dynamic range and use of separated protein spots in the detection technique. It also leads to a lack of sensitivity for less abundant proteins. Using current 2-D methods it is only possible to detect about 3,000 protein spots on an 18 x 20 cm^2 gel [73]. Yet, approximately 5,000-10,000 genes are expressed in a cell at any given time, resulting in the creation of at least 20,000-30,000 distinct proteins (due to alternative splicing and post-transitional modifications).

Another drawback of the gel approach is limitations of imaging and quantification systems which have led many to use manual examination to verify the accuracy of detected spots. This necessary verification process is a major bottleneck in efforts to automate such proteomic methods.

HPLC is a protein separation method most commonly used after protein digestion. In this approach, the proteins in a sample are primarily digested (cleaved into smaller peptides) using a protease such as trypsin. The chromatography portion of this method involves a separation method typically based on one of the following attributes [46]:

- Hydrophobicity: lacking attraction to water
- Strong cation exchange: net positive charge
- Strong anion exchange: net negative charge
- Size separation: size/molecular weight
- Special affinity: interaction with particular functional groups

Multidimensional liquid chromatography, or tandem liquid chromatography (LC), is the process of running a sample through two or more steps of LC and then separating the peptides based on multiple attributes. This creates a more refined subset of the original mixture of peptides. Multidimensional LC coupled with tandem MS (LC-LC-MS/MS) is a method used in the analysis of complex mixtures of peptides. This method is commonly known by the acronym Multi-Dimensional Protein Identification Technique, or MudPIT for short [74].

A significant advantage of MudPIT is its lack of 2-D electrophoresis, a time consuming and limited method as previously mentioned. The use of tandem LC increases the number of peptides that can be identified from fairly complex mixtures. As an example, in an analysis of a yeast cell lysate [75], 749 unique peptides (composing a total of 189 unique pre-digested proteins) were identified in a single MudPIT experiment, considerably more than even a single phase LC-MS/MS experiment. The most important advantage of tandem LC is the wide dynamic range of proteins that can be identified, eliminating the limitations presented by SDS-PAGE [76]. There are a number of works [75] where the analysis of a protein complex by 2-D electrophoresis yielded less identifiable proteins than the MudPIT approach. The only current drawback of MudPIT is the lack of sufficient bio-computing algorithms and visualization methods available to render the terabytes of data generated in a comprehendible format for scientists to examine. If these issues are solved, methods involving tandem liquid chromatography followed by MS techniques are set to significantly increase throughput and may ultimately replace 2-D electrophoresis in the long term.

B Protein Mass Spectrometry

Mass spectrometry (MS) is turning out to be one of the high growth areas in proteomics research in recent years. As shown in Figure 7, the field of mass spec in general has grown over 2 ½ times over the past decade in terms of PubMed related publications measured as discussed in “Technologies & Automation in Proteomics” section. This compares to a 1/3 increase in overall PubMed research article publications. Part of this growth is due to Mass spectrometry’s new applications in proteomic domains (as opposed to classical analytical chemistry-affiliated molecular studies) such as proteome mining, post-translational modifications, and protein-protein interactions. The immense amounts of data generated by MS based proteomics have paved the way for systematic identification of proteomes and intra-cellular dynamics. MS is also easily adaptable to high-throughput formats, a fact which has made it the method of choice for protein identification and characterization [73, 77]. An exhaustive review is not within the scope of this chapter, however the effort has been made to give an overview of the technology with biomedical applications of its use.

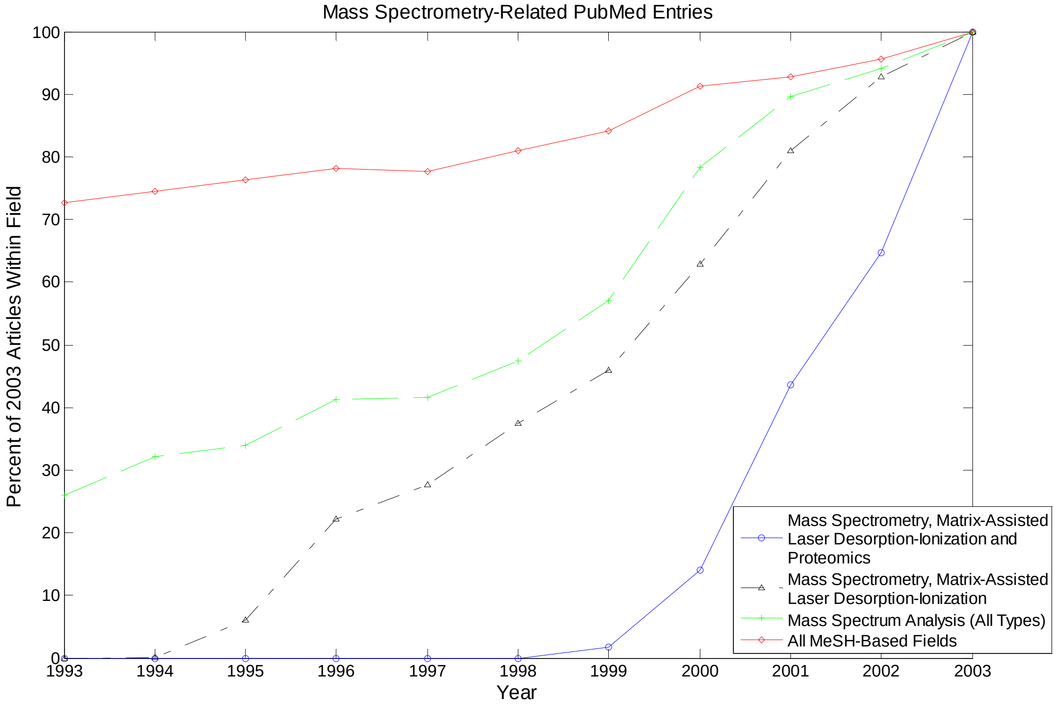


Figure 7: Mass spectrometry is growing at a much faster rate in terms of papers compared to the general PubMed database.

There are three main components in any MS machine: the source, mass analyzer, and detector. The source produces ions from the biological sample, the mass analyzer processes the ions in mass-to-charge (m/z) ratio dependent manner, and finally the detector detects the ions resolved by the mass analyzer. Fundamentally, MS converts the sample mixture into ions, analyzes them, and estimates their corresponding mass-to-charge ratios. The digestion of protein samples into small peptides (as described in the previous section) results in proteins being cleaved or cut between predictable amino acid locations. A database search of the masses is then carried out to decide which protein the sample peptide originated from. The described process demands high sensitivity, resolution and accuracy [78]. Sensitivity is required to measure masses on the order of femtomole (10-15) quantities with high resolution to distinguish between ions of the same m/z values.

Three prominent MS ionization methods used in proteomics are Electrospray Ionization (ESI), Matrix Assisted Laser Desorption/Ionization (MALDI) and Surface Enhanced Laser Desorption/Ionization (SELDI). In ESI mass spectrometry, a potential is applied to create a fine mist of charged droplets (including the dissolved peptide sample) that are subsequently dried and introduced into the mass analyzer. The solution used as input to the MS is often the output of HPLC (and includes digested proteins as well as the protease used to cleave them). In contrast to MALDI, ESI produces highly charged ions without fragmentation of the ions into the gas phase [73]. MALDI-MS is normally used to analyze relatively simple peptide mixtures, whereas integrated high performance liquid chromatography ESI systems (HPLC-ESI) are preferred for the analysis of complex samples.

The first step in the MALDI ionization source is the addition of the sample to a chemical matrix. The matrix includes photon absorbing molecules with a specific amount of chromophore, sensitive to light at a specific wavelength. The mixture is then placed on a small slide and allowed to dry. The dried mixture is a crystal lattice containing the desired sample to be analyzed. The crystal is then struck with a laser beam. The matrix molecules absorb the energy emitted by the laser, causing their temperature to increase. This excess heat causes the sample peptide to transform into gas phase [79]. Each peptide tends to (generally) pick up a single proton, creating a positive ion. This is significant since the m/z ratio is thus precisely the mass (z=1). This is in contrast to ESI where a peptide sample can pick up tens of protons, causing various peptides with the same mass to have differing m/z ratios. In any case, the ion then enters the mass analyzer where their m/z ratio-dependent behavior possible to differentiate between peptides present in the sample (e.g. see Equation 1 and accompanying text). SELDI is similar to MALDI; the ionization into the gas phase via photon absorption from a laser source remains the same. They differ in that SELDI sample plate surfaces are designed to react with peptide molecules with particular properties. Consequently peptides with similar physical and chemical attributes are retained, increasing their chance of becoming ionized and providing another layer of filtering (and decreasing required spectrum bandwidth) which helps in the identification of the peptides by a database search [80] or in creating diagnostically useful proteomics profiles.

SELDI has become increasingly popular since a landmark controversial study from the Liotta lab was first published in Lancet [81] involving diagnosis of ovarian cancer without actually identifying any proteins. As shown in Figure 7, the field of SELDI (indexed under MALDI in MeSH), measured in terms of papers, has grown very rapidly since being “introduced” as a category within MeSH in the 1990’s. The subset of MALDI/SELDI papers affiliated with proteomics has exhibited even faster growth.

As alluded to earlier, MS is also a clinical tool and has been used in numerous disease studies [4, 15, 82]. In an HIV study [83], MALDI was used to identify a family of proteins contributing to the CD8 antiviral factor, an important element in the pathology of AIDS. SELDI technology has also been applied to serum for cancer detection. Using machine learning techniques, early studies [4, 35] were the first to predict pathological states in their respective domains, such as ovarian cancer and preleukemia, solely using serum proteins. Rather than identifying proteins, such early studies yielded accurate diagnostic information in their respective fields based on the overall pattern of protein expression. In the case of ovarian cancer, the importance of early diagnosis is apparent in the high five year survival rate (95%) of patients with cancer limited to the ovary compared to a 35-40% five year survival rate for late stage patients [4]. SELDI has also been used in diagnosis of neurological diseases such as Alzheimer’s disease, Parkinson’s disease, multiple sclerosis, schizophrenia, and many others [82].

As with other processes in proteomics, MS sample preparation is undergoing automation and miniaturization. SELDI has been implemented to allow parallelization via multiple sample spots on arrays (see Figure 8). There are now “Lab on a CD” compact discs on which sample preparation procedures for peptide fingerprinting (or sequencing) by MALDI are miniaturized. A typical CD can prepare 96 protein samples simultaneously [71]. Systems are available that use robots to load a CD with reagents to purify proteins using micro-fabricated channels and centrifuge technology [55]. In the future, a fully automated high throughput protein analysis tool may be on a single chip. The Robot Automated Sample Preparation and

Analysis Pipeline for Proteomics (Raspap) system by Alterovitz, et.al. [35] successfully integrated hardware and software methods for SELDI-based proteomics analysis. The use of robotics and intelligent decision systems within this system will be discussed further in the ‘Integration’ section.

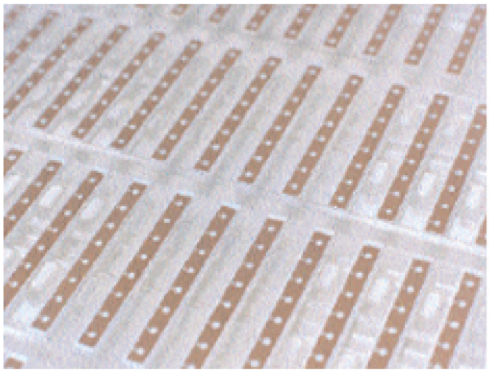


Figure 8. SELDI Array

With respect to mass spec technology, there are four basic types of mass analyzers currently used in proteomics research. These are the ion trap, time-of-flight (TOF), quadrupole time-of-flight (Q-TOF), and Fourier transform (FT-MS) ion analyzers. They are very different in design and performance and each with its own advantages. They can be used alone or put together in tandem to take advantage of the unique strengths of each [77].

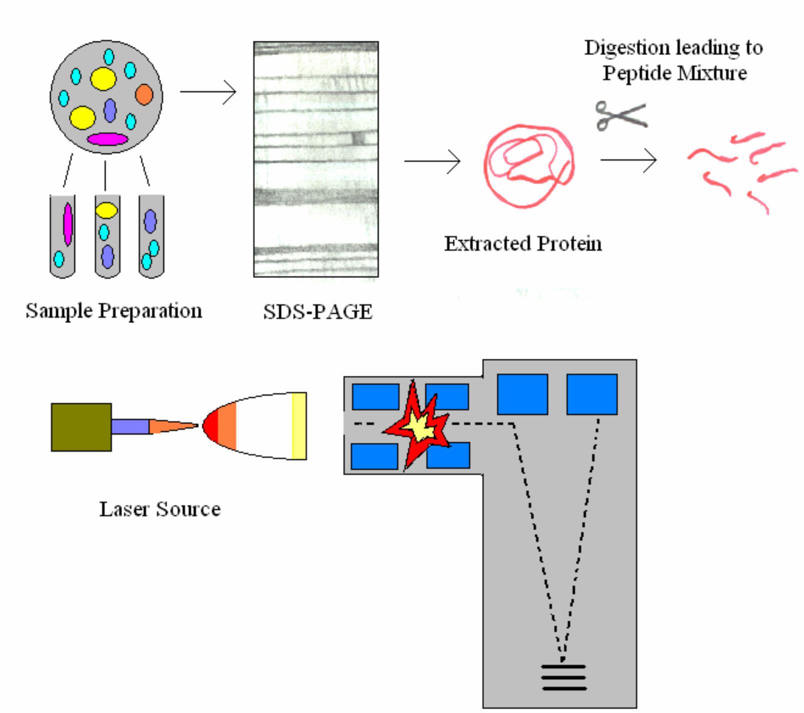


Figure 9. Steps Involved in tandem mass spectrometry

In the ion-trap analyzers, ions are first confined within a trap via electrically active electrodes on the top, bottom, and middle (via a ring electrode). The ion trap collects the ions for a certain time interval and then subjects them to MS or MS/MS analysis. Ion traps are robust, sensitive, and relatively inexpensive. They have produced a large percentage of the proteomics identification–related results reported in the literature [77]. The Fourier transform ion cyclotron resonance MS (FT-MS) is similar to an ion trap. This method however employs a magnetic field for detecting ions in the trap [84]. But in spite of the enormous potential in terms of measuring low abundance proteins, issues ranging from cost to operational complexity and low peptide-fragmentation efficiency have limited use of FT-MS instruments in proteomics research [79].

In TOF analyzers, time is measured for the gas-phase ions to travel from the ionization source to the detector, which is then related to the m/z ratio [85]. This analyzer is not as well suited for tandem MS (see below) and has the disadvantage of being dependent on sample quality for successful peptide identification [77]. A quadrupole mass analyzer is a variant of TOF that consists of four parallel metal rods that are arranged lengthwise. These can be manipulated to allow ions of a specific m/z ratio to pass between them for detection. The TOF analyzer is typically paired with MALDI (MALDI-TOF) or SELDI (SELDI-TOF) where as the quadrupole and Fourier transform methods use ESI sources. The equation governing TOF analyzers with some common values (e.g. for PBS II SELDI-TOF, Ciphergen, Fremont, CA) is shown below.

$$\frac{m/z}{U} = a(t - t_0)^2 + b$$

Equation 1: Physics -based formula for mass-to-charge ratio
Where:

t = time of flight (μs)
m = mass (Da)
z = charge
U = 20,000 Volts
a = 0.272, b = 0, t_0 = 0.0038 are constants

An overview of tandem MS (MS/MS) is shown in Figure 9. First, peptide ions generated from an ESI source are separated based on the m/z ratio. In the second round, a single m/z is chosen and is subject to Collision Induced Dissociation (CID) [86]. This process induces fragmentation of the peptide into fragment ions, which are then analyzed on the basis of their m/z. The resultant tandem spectra of amino acid composition can be searched against protein databases to identify the protein [87]. Matches from at least three to six peptides derived from the same protein are typically required to positively identify a protein [88]. Tandem MS also provides information about the nature and location of peptide modifications. The extent and comprehensiveness of the available databases are extremely crucial as database-searching strategies can be applied only if the protein sequence exists in the database. Sequest, developed at the University of Washington [89], is the most widely used tool for searching protein databases [90]. Sequest, discussed further in the next section, is ideal for high-throughput proteomics as it automatically extracts and searches the MS/MS data against a protein database [91].

It must be mentioned that although MS is a sensitive method for identifying proteins, there are quantitative shortcomings [92]. The intensity of a peptide peak depends linearly on the concentration of the peptide. However, different peptides have different propensities for ionization. Thus, two peptides present in equal amounts may show substantially different intensities in the mass spectra. This problem has been addressed by modifying one of the sample types with a stable isotope (e.g. the cancer samples) while leaving the other unchanged (e.g. the control samples). This changes the molecular weight of the isotope-based samples, but not the mass spectrometer's behavior in terms of the peak intensities. Quantitative differences are then determined directly as the difference in peak area between the two peptides in the mixed sample [53].

C Database Search Algorithms for MS and MS/MS Spectra

Following MS or MS/MS processing, a database search can be carried out to try to identify proteins of interest. One such method, known as peptide mass fingerprinting, involves identification of a protein given peptide MS information. Protein identification with tandem MS (MS/MS) data and the Sequest algorithm is a second approach.

Following application of analytical protein separation methods such as 2-D electrophoresis, digestion of the excised proteins, and MS (e.g. MALDI-TOF) on the resulting peptides, one obtains a set of m/z ratios of the peptides present in the sample. One goal in proteomics is to determine the protein identities with high certainty. The success of the identification process is dependent on the quality of MS data, the accuracy of the database, and the power of the search algorithm used [93].

In a typical identification algorithm, a database of known proteins is set up (e.g. using SWIS-Prot, OWL, and/or NCBInr). A protease is specified and used for virtual (i.e. *in silico*) protein digestion to yield a master peptide list with corresponding masses. Matches are made between peptide masses obtained from MS and the peptide master list. If several of these peptides uniquely match the same protein, then the unknown sample protein can be identified. The process is also applicable if there are multiple proteins. In that case, however, there is more room allowed for error and a scoring system is typically used to rank the fidelity of each match. Most scoring systems assign higher scores to those proteins with the greatest number of peptide matches. This tends to give bigger proteins a higher score, simply because they yield more peptides upon digestion [46]. Some probability based scoring algorithms have emerged [94]. One such algorithm is ProFound [95].

ProFound ranks protein candidates using a Bayesian-based algorithm, taking into account individual properties of proteins in the database as well as other information relevant to the experiment. The algorithm assumes that the candidate protein is contained in the database and that all the detected peptide ions come from the protein under consideration. A hit is a match between a measured mass and a calculated theoretical peptide mass given an accuracy range. The ranking is directly proportional to $P(k \mid D,I)$, namely the probability for each hypothesis k given data D and background information I. This score is calculated as shown in Equation 2 below.

$$P(k \mid I,D) \propto P(k \mid I)\frac{(N-r)}{N!}\prod_{j=1}^{r}\left(\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{\sigma_j}\,\sum_{i=1}^{g_j} e^{-\frac{(m_j-m_{ij})^2}{2\sigma_j^2}}\right) F_{pattern}$$

$$\sum_{k \in database} P(k \mid I,D) = 1$$

Equation 2: ProFound Bayesian-based algorithm

In Equation 2, the variables are defined as follows:

K : hypothesis that: protein k is the protein being analyzed

D : the experimental data

I : available background information about the protein (species of origin, enzyme cleavage chemistry, approximate molecular mass, previous experiments, etc.)

N : the theoretical number of peptides generated by fragmentation of protein k given a protease.

r : the number of hits

m_{max} - m_{min} : the range of measured peptide masses

m_i: the measured mass of the ith hit

g_i : the number of theoretical peptides that match m_i

m_{ij} : the calculated mass of the jth peptide in the ith hit

σ_I : the standard deviation of the mass measurement at mass m_i

$F_{pattern}$: an empirical coefficient

It has been shown that the above algorithm is superior in performance to its predecessors (which not employ such probabilistic reasoning) [95].

Peptides in the human body are composed of a chain of the 20 amino acids available in humans. These amino acids are represented by a letter in the literature and have various masses, see Table 1.

Protein identification using tandem MS (MS/MS) experiments employs different algorithms, taking advantage of the second MS spectrum. A peptide is a sequence of amino acids and hence its mass is the equal to the sum of the masses of the amino acids that compose it. However, since the order of the amino acids is important in determining a peptide’s structure/function, permutations of a sequence of amino acids may yield different peptides with the same masses. In addition, some amino acids (e.g. isoleucine and leucine) or modified amino acids may have the equivalent masses (either due to identical masses or limits in a measuring instrument’s precision). In MS/MS, data peptides of a specific mass are selected and subject to collision induced dissociation, resulting in two sequences of amino acids referred to as fragments. As an example, GVAGNEGAL is a peptide which can be fragmented into GVAG and NEGAL ions. If all GVAGNEGAL peptides were fragmented into GVAG and NEGAL ions, it would not be possible to recover the peptide’s sequence. However various GVAGNEGAL peptides will break at different points along the sequence. This is crucial to MS/MS since then the fragments can be pieced together in the correct order. The resulting spectra can then be analyzed to obtain the sequence (see Figure 9).

Table 1: Amino acids and corresponding molecular weights

Amino Acid	Symbol	Average molecular weight (da)
Alanine	A	71.0788
Arginine	R	156.1876
Asparagine	N	114.1039
Aspartic Acid	D	115.0886
Cysteine	C	103.1448
Glutamine	Q	128.1308
Glutamic Acid	E	129.1155
Glycine	G	57.0520
Histidine	H	137.1412
Isoleucine	I	113.1595
Leucine	L	113.1595
Lysine	K	128.1742
Methionine	M	131.1986
Phenylalanine	F	147.1766
Proline	P	97.1167
Serine	S	87.0782
Threonine	T	101.1051
Tryptophan	W	186.2133
Tyrosine	Y	163.1760
Valine	V	99.1326

There are two approaches to resolving MS/MS spectra into a peptide sequence. The de novo method involves manual analysis by an experienced scientist using the above table to generate a predicted peptide sequence. Needless to say, this manual approach has not proven to be the best method for high throughput applications. The de novo method is usually followed by a search of a virtually digested protein database, similar to peptide mass fingerprinting, to identify the protein the peptide originated from. Algorithms have been developed to resolve MS/MS spectra into peptide sequences. The Sequest algorithm is the most commonly known for such analysis [96, 97]. A description of the algorithm follows.

Sequest generates identifications using two pieces of information: the m/z ratio of the peptide before fragmentation (obtained from the first MS step) and the MS/MS spectrum. The m/z value of a peptide being analyzed with the peptide master list generated from a virtually digested protein database (as in peptide mass fingerprinting). A set of peptides within a specified mass range similar to the peptide m/z are chosen. These virtual peptides are processed to produce theoretical or model MS/MS spectra. The actual MS/MS spectrum is compared to the every model spectrum and a cross correlation score (XCorr) is given to each comparison. The XCorr value is dependent on the quality of the tandem mass spectrum and the quality of its fit to the model spectrum. Sequest creates a model MS/MS spectrum based on elementary knowledge of how peptides fragment in the collision induced dissociation process. The XCorr value generated during the analysis is not an absolute measure of spectral quality and closeness of fit to the model spectrum. That is, the algorithm will identify the best matches between the model and actual spectra regardless of the quality of the fit. Thus, the same XCorr value for one peptide may not mirror a similar closeness of fit for another peptide with the same score.

Scoring Algorithm for spectral analysis (SALSA) is a feature extraction algorithm designed to identify and score particular features in MS/MS spectra. SALSA aims at solving problems in identifying a subset of the sample proteins with specific characteristics. Examples of such scenarios are: the detection of peptides with a particular amino acid sequence (motifs) and the identification of protein modifications such as phosphorylation. More specifics regarding SALSA can be found in several other sources [98-100].

ProFound, Sequest and SALSA present the capability to rapidly render data into useful tangible information. These algorithms, when coupled with automated sample preparation and MS techniques such as HPLC-MS/MS, are enabling identification of hundreds of proteins.

IV Statistical and Machine Learning Methods

Statistical learning and data mining techniques make it possible to do automated data mining even as biological databases grow exponentially. Techniques such as artificial neural networks (ANN) [101], support vector machines (SVM) [102], genetic algorithms (GA) [103], and statistical regression techniques provide tools for supervised learning when training data is available (with appropriate class labels that help to ‘supervise’ the algorithm and guide its learning). When the class labels are not available (i.e. unsupervised learning), various clustering techniques can be used to find structure in the data. Numerous nonapplication-specific algorithms exist such as K-means clustering [104], principal component analysis (PCA) [105], pairwise hierarchical clustering [106], and Bayesian techniques [107].

Bayesian clustering algorithms have been used with success in both supervised and unsupervised learning. Examples of Bayesian strategies for genomic micorarray data include CAGED [108] and Botstein’s approach [109]. By using the ARPA approach as discussed in the ‘Integration’ section, data from proteomics can be studied in a similar fashion. One such work involves pathologic detection via a Naïve Bayesian Classifier based on SELDI data [35]. Here, the abundance of various biomarker peaks (proxies to protein abundance) is used to predict whether or not a patient has a particular pathological condition. In the following subsections, several of these methods will be described in more detail including Bayesian Learning, Support Vector Machines, and Principal Component Analysis.

A Bayesian Methods

Bayesian methodology facilitates inclusion of *a priori* information (e.g. from an expert) in order to facilitate inference on a dataset. It helps characterize the parameters’ conditional probability given *a priori* information by looking at the parameter vector as a probability distribution that can be conditioned upon. The classical example is the flipping of a coin. Whether an object landing on the ground is a fair coin or a magician’s biased coin can influence the probability that one expects heads to come up- before the coin is even tossed.

While classical statistics would glean this information from multiple tosses, a Bayesian approach would incorporate this information by calculating the prior density:

P(parameter vector | *a priori* information)

With limited examples, this approach would likely perform better than the classical statistical approach. As the number of examples increase, the Bayesian results often approach those of classical methods. In proteomics, the data is limited due to cost considerations and the novelty of the field. Thus, the Bayesian approach will be suitable to help capture the structure of the data with the limited number of available cases.

Bayesian probabilistic assumptions and relationships can be visualized through graphical models (known as Bayesian Networks). A Bayesian Network is essentially a graphical representation of probabilistic dependencies. Let G={V, E} be a directed acyclic graph (DAG) with V representing vertices and E being a vector of edges. In such a graph, the vertices typically encode the variables and directed edges imply probabilistic dependence. These dependencies help reduce the number of terms in the joint probability and hence reduce the amount of computation needed for inference.

A Naïve Bayesian Classifier (NBC), is shown in Figure 7. Here, the information encoded is that the attributes X_1 to X_N are conditionally independent given their mutually exclusive classes Y (MDS or Control). In other words, $(X_1 \ldots X_N)$ are $\perp$ | Y. In this case, there are N attributes- where N is the number of biomarker (or protein) peaks.

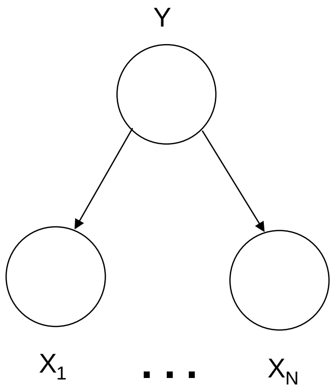


Figure 10: Naïve Bayesian Classifier: Directed Graph with Conditional Independence Assumption

Through application of Bayes' Rule, marginalization, and conditional independence assumptions, Bayesian inference can be used to compute the posterior probability distribution of the class variable given a set of sample attributes, such as the posterior probability distribution of a sample being a tumor given that some proteins are present [110].

An advantage of the Bayesian approach is that it can be used to capture expert knowledge while at the same time incorporating data-based information. By taking into account accurate application-specific information and data dependencies, better clustering has been shown than through generic clustering algorithms [111, 112].

B Support Vector Machines

SVM has also been used for many medical applications including microarray-based gene classification [113] to blood maturation categorization [114]. Recently, SVM has been applied to bioinformatics and proteomics as well [35, 115-117].

SVM is a supervised learning technique that can be viewed as a Tikhonov Regularization problem with a hinge loss function. That is, it can be expressed as:

$$\min_{f\in H} C\sum_{i=1}^{n} L(y_i, f(x_i)) + \frac{1}{2}\|f\|_K^2$$

where the loss function is: $L(y, f(x)) = \max[1 - y * f(x), 0]$. Here H is the Hilbert space and K is the Reproducing Kernel Hilbert Space (RKHS) used to define the norm. C is the regularization constant. X represents the biomarker peak value(s). Y is the actual MDS diagnosis whereas f(x) is the predicted MDS status.

To solve this regularization problem, one can rewrite it as a constrained quadratic programming problem with Lagrange multipliers:

$$\max_{\alpha\in\Re n^n} \sum_{j=1}^{n} \alpha_i - \frac{1}{2}\overline{\alpha}^T \overline{\overline{Q}}\,\overline{\alpha}$$

where: $Q = \overline{y}K(x_i, x_j)\overline{y}^T$ and K(x_i, x_j) is the kernel function

with constraint $\overline{y}$ is orthogonal to $\overline{\alpha}$ (i.e. $\overline{y^T}\,\overline{\alpha} = 0$) where:

$$0 \le \alpha_i \le C \qquad i = 1...n$$

This can be solved using normal quadratic programming techniques (e.g. as implemented in Matlab).

C Principal Component Analysis

PCA has been used successfully previously to discover relevant components within medical datasets for analysis, clustering, and compression purposes[118]. In PCA, each principal component is an eigenvector consisting of weighted parameters (protein biomarker peaks in this case). The importance of a given principal component in terms of explaining the data variance is represented via eigenvalues which are determined as explained below.

The principal components are found as follows. First, the covariance matrix (e.g. *E*) of the data matrix *L* is calculated. Next, the eigenvalues and eigenvectors of *E* are found. The eigenvectors are sorted (to form a matrix, *Q*) so that they are in descending order based on the eigenvalues. Next, the first n eigenvectors in *Q* (with largest eigenvalues) are selected based on a scree plot [105] to form matrix *Z*. A scree plot involves plotting the eigenvalues magnitudes for each eigenvector and comparing the difference between them in order to select those above a noise baseline (a lower slope magnitude is typically prevalent at less significant components). Within each principal component, the eigenvlaue magnitudes are ranked and the corresponding biomarker peaks can then be determined.

V Case Studies

This section illustrates the concepts and notions of statistical analysis of proteomics data described in the previous section using some actual case studies. For more information on the models used, BAP (Bioinformatics Analysis Pipeline), and ARPA (Analysis and Robot Pipelined Automation) 2, please see
http://www.chip.org/proteomics/pub/foris2004/index.html.

A Challenges in Statistical Models for Mass Spectrometry

While some of the fundamental physics of mass spectrometry technologies has been worked out, not all of details are known. For example, the models for the mechanism of ionization have not proved sufficient in predicting spectrums accurately (which influences the m/z ratio). Also, concentration cannot be used solely to predict the intensity of the associated peaks as numerous other variables are involved such as solution composition and mass spec behavior [119]. Yet, even if the intensity can be associated with one peptide mass, there are still challenges in associating this with a unique peptide. While MS/MS techniques typically use Sequest-like methods, SELDI/MALDI techniques cannot (due to the lack of the second MS signal information). As a result, mostly proteomic profiles have been reported via these techniques rather than an in-depth analysis of myriad proteins.

The problem is that many proteins have similar masses, so it is hard to uniquely identify a protein based solely on mass (even assuming single ionic charge, z=1). Figure 11 plots proteins present per mass unit (Da) in the mass range of 700 - 12,000 Da, the same range used the most recent high resolution SELDI mass spec instrument in terms of dynamic range [120]. Even with an ideal intra-machine mass drift of approximately 100 ppm (parts per million), this SELDI-based instrument cannot be more accurate than +/-1.2 Daltons (Da) at 12000 Da. In the aforementioned study, bins of 400 ppm were used after analysis was done to estimate the best window to accommodate inter and intra assay variance and drift. Older generation instruments had margins of error two orders of magnitude higher than this. Other constraints to mass spectrometry-based proteomics include the probabilities associated with finding particular peptides extracellularly- and within the type of tissue or body fluid sampled.

To generate the data visualized in Figure 11, Entrez was searched for all proteins with molecular weight within 1 Da windows along the 700 - 12,000 Da mass range. This search included cleavage products of proteins and protein precursors (based on annotation features) for a more accurate picture of potential proteins that might be found upon mass spec. This yielded 1,043,613 protein entries in this range, such that about 1/5 of all Entrez protein entries can be examined within this mass spec technology range. Since most SELDI studies have focused on human proteins, this subset was focused on next, yielding 46,843 protein entries, again around 1/5 of the total number of human entries.

Since the Entrez Protein database contains redundant entries, SeqHound [121] was used to determine the non-redundant entries via remote Java API (Application Protocol Interface) calls implemented in Matlab. This reduced the number of human proteins to 36,682. Finally, each database sequence was examined for similarity directly, further reducing the number of non-redundant proteins to 36,024. Thus, it is expected that there would be around 3 proteins per mass unit (Da) given the 700 - 12,000 Da range. While the average number of proteins per mass unit is 3.19, the distribution across the mass range clearly shows a pattern rather than uniformity- with peaks at approximately 2,300 Da and 11,600 Da among others.

There are some unusually high counts and outliers at the point labeled 'A' in the Figure 12. Exploration of these suggested the commonality was in slightly different proteins dealing with T-cell receptor beta/delta chains. For the points demarcated 'B,' different forms of immunoglobulin heavy/light chain variable regions were seen. Since immunology demands a great diversity of T-cell receptor chains and immunoglobulin variable regions, knowing the masses where these molecules are concentrated could potentially help in tuning protein identification algorithms as well as yield insights into the relevant biology. In fact, T-cell receptor loci and immunoglobulin loci both have gene segments with variable regions that are rearranged by exactly the same enzymes [122].

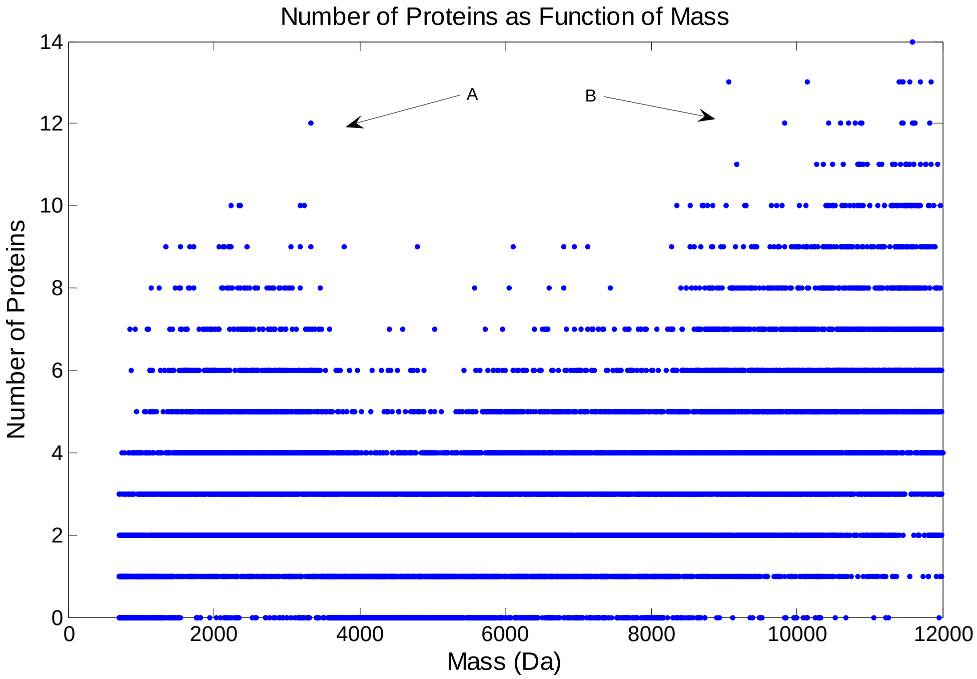


Figure 11: The number of human proteins per mass unit for various masses is rarely unique

It would be useful to be able to model the number of proteins per mass unit for probabilistic calculations. Often, the Poisson distribution is used to model situations involving counts or arrivals during an interval of time [123]. In this case, at each mass unit, an average number of proteins are expected to 'arrive' during this interval. However, the Poisson distribution has only one parameter (λ) which is equal to the mean and variance. Yet, since the variance is 4.82 (compared to 3.19 for mean), the Poisson model is not a good fit in this case (see Figure 12). In such cases, the negative binomial distribution can be used as it has two parameters (p and r) [124] and is a superset of the Poisson distribution (approaches it in the limit $r \to \infty$). It has been used commonly in modeling count data as well. For example, one common application is modeling daily road accidents at certain highway locations [125]. As there can be high variances in this scenario (daily accidents dependent on the day's weather conditions, etc), negative binomial models have been used instead of Poisson in such cases. The negative binomial distribution is commonly defined as:

$$f(x \mid r, p) = \left(\frac{r+x-1}{x} \right) p^{r} (1-p)^{x}$$

Equation 3: Negative binomial distribution for $r \in \square^{+}$

However, when parameter r is not restricted in integer values, the more general expression becomes [124]:

$$f(x \mid r, p) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^{r} (1-p)^{x}$$

The gamma function from above is defined as:

$$\Gamma(x) = \int_{0}^{\infty} t^{x-1} e^{-t} dt$$

The gamma distribution (not to be interchanged with the gamma function) is useful in measuring failure times and is a superset of the exponential distribution since it allows for an additional dependence on the 'age' of the item [123]. In this case, the failure times would be the count of proteins that are confined to a certain mass before the next mass window starts.

The above models assume independent, identically distributed (IID) counts of proteins per mass unit (not perfectly satisfied as can be seen in Figure 11). Each of the aforementioned distributions' parameters were then estimated via maximum likelihood estimation (MLE). The resulting models are shown in Figure 12. The negative binomial model (with MLE estimates of r=6.19 and p=0.660) was the best fit by several measures. A Monte Carlo simulation of the Wilcoxon rank sum test at the 5% significance level was performed to test concordance of the models with the data. The negative binomial distribution was the best fitting model. It also had the best log likelihood score at: -2.39x10^{4}.

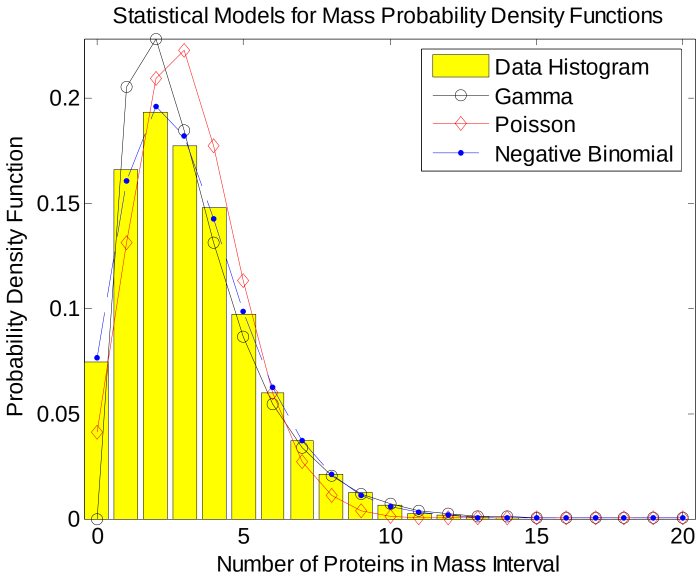


Figure 12: The probability density function of the data appears to follow a negative binomial distribution

B SELDI/ESI Proteomics Pipeline Automation and Integration

One of the major challenges in bioinformatics in general is integration and analysis of high throughput/disparate data sources. Currently, automated systems have been built for SELDI and ESI separately. However, as discussed in the previous section, each mass spec technology has its own advantages. As costs decrease, biocomputing centers and labs are able to buy or share mass specs that use different technologies. In this section, we describe a new automated system ARPA 2 (Analysis and Robot Pipelined Automation) that extends Raspap's [35] analysis to local ESI (currently implemented at Harvard Partners Center for Genetics and Genomics) and SELDI-type (as implemented in Raspap) data as well as remote Open Proteomic Database (OPD) data. In doing so, we will do first the machine learning-based analysis comparing head/neck versus cervical cancer, using ESI mass spec data from Mark Carlson [126]. The overall structure of ARPA 2 and its inputs is schematically shown in Figure 14 below.

Cancers of the cervix, or the neck of the uterus, have a number of similarities to head/neck cancers. Like cervical cancers, head/neck cancers are often squamous cell carcinomas. They also share histology, epidemiology, and exposure-related characteristics. For example, human papillomavirus (HPV) exposure plus smoking have been shown to work together as cofactors linked to both cancers [127]. Thus, it would be interesting to analyze the differences/similarities in the proteins found in both conditions. The dataset examined here included 22 samples, divided evenly between the two conditions. Human cell line SqCC (squamous carcinoma cells) were used to model head and neck cancer while SiHa human cell lines [128] were used to prototype cervical cancer. Cell lysate samples were run by the Carlson group on the Dexa XP Plus ESI-Ion Trap mass spec (Thermo Electron Corporation, Waltham, MA, USA) with further protocols and specifications available online [126, 129].

Using the raw MS/MS data, we used the Sequest algorithm, as implemented in Bioworks (Thermo Electron Corporation) for sample protein identification and estimated peak area information per protein. We performed further processing on the generated CSV (comma separated value) file suitable for the BAP (Bioinformatics Analysis Pipeline) component of ARPA 2 (See Figure 13).

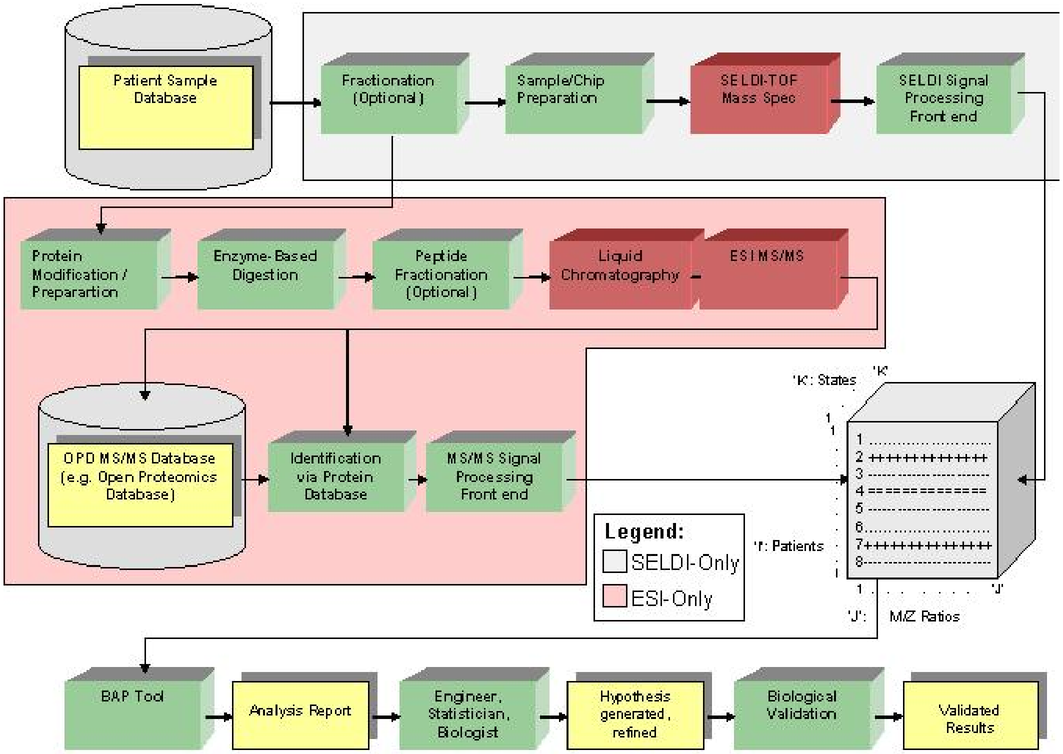

Patient Sample Database
Fractionation (Optional)
Sample/Chip Preparation
SELDI-TOF Mass Spec
SELDI Signal Processing Front end
Protein Modification / Preparartion
Enzyme-Based Digestion
Peptide Fractionation (Optional)
Liquid Chromatography
ESI MS/MS
OPD MS/MS Database (e.g. Open Proteomics Database)
Identification via Protein Database
MS/MS Signal Processing Front end
'K': States
'K'
'I': Patients
'J': M/Z Ratios
1
'J'
Legend:
SELDI-Only
ESI-Only
BAP Tool
Analysis Report
Engineer, Statistician, Biologist
Hypothesis generated, refined
Biological Validation
Validated Results

Figure 13: Schematic showing how local SELDI, ESI, and remote OPD data can be integrated via common format for analysis

The Sequest XCorr value was used for monitoring the intra- and inter-sample quality of the protein identifications. Using BAP, 70% of the data was reserved for training and the remainder for testing using the following models: k-nearest neighbor, support vector machine (SVM), and logistic regression. The k-nearest neighbor classifier performed best with an accuracy of 85.7%, specificity of 100%, and sensitivity of 75%. The root mean square error (RMSE) was 0.378. Optimizing for the k parameter led to k=1 as the best value. Essentially, the nearest training set point to each point in the test set was used in making the predictor. Since the dataset was small, this was likely the best since more complicated models with higher k values could lead to overfitting on the training set.

Since the k-nearest neighbors algorithm can be black box in terms of explaining the rationale behind the decision process, the more intuitive decision table algorithm [130] was used to derive simple rules based on the protein peak levels. This is a similar type of approach taken previously, where a tree-based method was used to achieve higher predictive accuracy (85%) and specificity (80%) for preleukemia compared to previous work- while using just five proteins (instead of hundreds) and three rules [115]. Here, this was done using a Java interface to Weka [131], something that can be done from BAP/Matlab via the integrated Java Virtual Machine (JVM). Creating the decision table using cross validation led to a set of just two rules (see Figure 14). Though sensitivity (72.7%) and specificity (63.6%) were lower than the original k-nearest neighbor method, it also only used two proteins instead of 430.

If (GenBank_ID_4885431 < 33650000 peak area) and (GenBank_ID_7669492 > 279750000 peak area), then declare 'head/neck.'
Otherwise, declare 'cervical.'

Figure 14: Decision table rules

Interestingly, both of these proteins have been associated with cancers in the head/neck. Glyceraldehyde-3-phosphate dehydrogenase (GenBank ID [132]: 7669492), a 335 amino acid protein, has been linked to thymona, a cancer in the thymus of the neck. It has also been connected to apoptosis (cell death) in human breast cancer cell lines [133, 134]. Heat shock 70kDa protein 1B (GenBank ID [132]: 4885431), a 641 amino acid protein, has been associated with nasopharyngeal carcinoma [135].

VI Conclusion

In this chapter, the applications of robotics and intelligent decision systems within proteomic were introduced along with novel work to illustrate the research issues involved. The contribution of genomics in understanding proteomes is invaluable. However, the greatest complexity lies in the diversity of the full set of protein products and interactions after gene transcription is already complete. As the number of proteins being cataloged in databases continues to grows exponential (see Figure 15) while the estimates of the number of genes in humans and other organisms is actually declining [2], the opportunities for proteomics to make use of this information grows. As such, novel statistical and engineering-based methods will be needed to analyze this information.

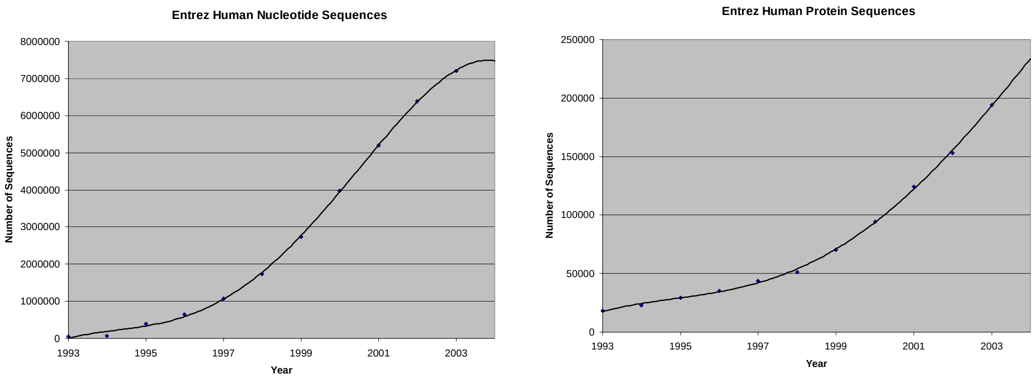


Figure 15: While the number of genetic sequences in Entrez is starting to saturate, the proteins being cataloged in Entrez is still growing exponentially each year.

Proteins' abundance, miniature size, and dynamic nature have made them difficult to explore. On the other hand, these features also make proteins the perfect complex system for engineering-based analysis. Precision and throughput are key parameters for technologies addressing these issues. Accurate sensors and signal detection methods are needed to indicate protein abundance and interaction. High throughput robotic systems will significantly increase efficiency and reduce the potential for error in sample preparation and processing. Intelligent decision making systems for image analysis (e.g. for gels), feature extraction, and other machine learning techniques will reduce the burden on scientist in analyzing experimental results and make whole-organism proteome-based experiments a reality.

As seen in the case studies, new research in proteomics needs to build on and leverage a technology's strengths while at the same time integrating other data sources to make the best possible use of available information. Both engineering and scientific expertise are needed in evaluating the conclusions. For example, determining the validity and relevance of proteins requires biological expertise while the design of a protein chip or statistical algorithm requires a different technical background. Thus, making good use of information gleaned during such experiments requires innovative approaches ranging from constructing accurate cellular models to better experimental hypotheses. In this new era, proteomics is not just validating hypotheses, but also generating new ones.

The immense clinical potential and promise of proteomics has also begun to burgeon in disease diagnosis, prognosis, and treatment. HIV, various neoplastic entities (i.e. cancer), immunological disorders, and many more pathological ailments are targets for clinical proteomics. In short, proteomics is set to change the way people view cellular function, disease, and humanity itself.

Acknowledgement

The authors would like to acknowledge Gabrielle Whitcomb (who along with Ehsan Afkhami) drew figure illustrations for this chapter. This work was supported in part by a fellowship from the Whitaker Foundation.

References

[1] Lander, E.S., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. 409(6822): p. 860-921.

[2] Human Genome Sequencing Consortium, I., Finishing the euchromatic sequence of the human genome. *Nature*, 2004. 431(7011): p. 931-945.

[3] Wilkins, M.R., From proteins to proteomes: large scale protein identification by two dimensional electrophoresis and amino acid analysis. *Biotechnology*, 1996. 14: p. 61-65.

[4] Petricoin, E.F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A., Clinical proteomics: translating benchside promise into bedside reality. *Nature Rev. Drug Discovery*, 2002. 1: p. 683-695.

[5] Wadsworth, J.T., et al., Serum protein profiles to identify head and neck cancer. *Clin Cancer Res*, 2004. 10(5): p. 1625-32.

[6] Carrette, O., et al., A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease. *Proteomics*, 2003. 3(8): p. 1486-94.

[7] Dare, T.O., et al., Application of surface-enhanced laser desorption/ionization technology to the detection and identification of urinary parvalbumin-alpha: a biomarker of compound-induced skeletal muscle toxicity in the rat. *Electrophoresis*, 2002. 23(18): p. 3241-51.

[8] Mukhopadhyay, T.K., et al., Rapid characterisation of outer membrane proteins in Neisseria lactamica by surface enhanced laser desorption and ionisation - time of flight mass spectroscopy for use in a meningococcal vaccine. *Biotechnol Appl Biochem*, 2004.

[9] Gravett, M.G., et al., Diagnosis of intra-amniotic infection by proteomic profiling and identification of novel biomarkers. *Jama*, 2004. 292(4): p. 462-9.

[10] Xiao, Z., et al., Serum proteomic profiles suggest celecoxib-modulated targets and response predictors. *Cancer Res*, 2004. 64(8): p. 2904-9.

[11] Boot, R.G., et al., Marked elevation of the chemokine CCL18/PARC in Gaucher disease: a novel surrogate marker for assessing therapeutic intervention. *Blood*, 2004. 103(1): p. 33-9.

[12] Anderson, N.L., et al., The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics*, 2004. 3(4): p. 311-26.

[13] Davis, M.T., et al., Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. *Proteomics*, 2001. 1(1): p. 108-17.

[14] Spahr, C.S., et al., Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. *Proteomics*, 2001. 1(1): p. 93-107.

[15] Hanash, S., Disease proteomics. *Nature*, 2003. 422: p. 226-32.

[16] D.D. Shoemaker P.S. Linsley Recent developments in DNA microarrays. *Current Opinion in Microbiology*, 2002. 5: p. 334-7.

[17] Templin, M.F., et al., Protein microarrays and multiplexed sandwich immunoassays: what beats the beads? *Comb Chem High Throughput Screen*, 2004. 7(3): p. 223-9.

[18] Nielsen, U.B. and B.H. Geierstanger, Multiplexed sandwich assays in microarray format. *J Immunol Methods*, 2004. 290(1-2): p. 107-20.

[19] Xu, Q. and K.S. Lam, Protein and chemical microarrays-powerful tools for proteomics. *Journal of Biomedicine & Biotechnology*, 2003. 2003(5): p. 257-66.

[20] Hosokawa, Y., et al. Fabrication and application of protein crystal microarrays. in Bioinspired Nanoscale Hybrid Systems. Symposium, 2-4 Dec. 2002. 2003. Boston, MA, USA: Mater. Res. Soc.

[21] Smith, J.T. and W.M. Reichert. The optimization of quill-pin printed protein and DNA microarrays. *in Conference Proceedings. Second Joint EMBS-BMES Conference 2002 24th Annual International Conference of the Engineering in Medicine and Biology Society*. *Annual Fall Meeting of the Biomedical Engineering Society*, 23-26 Oct. 2002. 2002. Houston, TX, USA: IEEE.

[22] Gosalia, D.N. and S.L. Diamond. High throughput screening using enzyme assay microarrays. in *Conference Proceedings. Second Joint EMBS-BMES Conference 2002 24th Annual International Conference of the Engineering in Medicine and Biology Society. Annual Fall Meeting of the Biomedical Engineering Society, 23-26 Oct. 2002*. 2002. Houston, TX, USA: IEEE.

[23] Lee, K.-N., et al., Micromirror array for protein micro array fabrication. *Journal of Micromechanics and Microengineering*, 2003. 13(3): p. 474-81.

[24] Jin, G., et al. Immune-microassay with optical proteinchip for protein detection. in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 17-21 Sept. 2003*. 2003. Cancun, Mexico: IEEE.

[25] Najmabadi, P., A.A. Goldenberg, and A. Emili, Conceptual design for an automated high-throughput magnetic protein complex purification workcell. *JALA*, 2003. 8(6): p. 101-6.

[26] Muthusubramaniam, L., et al. Automating crystallization of membrane proteins by robot with soft coordinate measuring. in *2004 IEEE International Conference on Robotics and Automation, 26 April-1 May 2004*. 2004. New Orleans, LA, USA: IEEE.

[27] Kazerounian, K., From mechanisms and robotics to protein conformation and drug design. Transactions of the ASME. *Journal of Mechanical Design*, 2004. 126(1): p. 40-5.

[28] Lee, W.C. and Y.-H. Cho. Nanomechanical protein detectors using electrothermal nano-gap actuators. in *17th IEEE International Conference on Micro Electro Mechanical Systems. Maastricht MEMS 2004 Technical Digest, 25-29 Jan. 2004*. 2004. Maastricht, Netherlands: IEEE.

[29] Pan, Y.V., et al. A precision technology for controlling protein adsorption and cell adhesion in bioMEMS. in *Technical Digest. MEMS 2001. 14th IEEE International Conference on Micro Electro Mechanical Systems, 21-25 Jan. 2001*. 2001. Interlaken, Switzerland: IEEE.

[30] Song, G. and N.M. Amato. A motion planning approach to folding: from paper craft to protein folding. in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation, 21-26 May 2001*. 2001. Seoul, South Korea: IEEE.

[31] Song, G. and N.M. Amato, A motion-planning approach to folding: from paper craft to protein folding. *IEEE Transactions on Robotics and Automation*, 2004. 20(1): p. 60-71.

[32] Bertone, P. and M. Gerstein, Integrative data mining: the new direction in bioinformatics. *IEEE Engineering in Medicine and Biology Magazine*, 2001. 20(4): p. 33-40.

[33] Kohlbacher, O. and K. Reinert, Differential analysis in proteomics: experimental methods, algorithmic challenges. *IT-Information Technology*, 2004. 46(1): p. 31-8.

[34] Hai-ting, Z., Machine learning and bioinformatics. *Information and Control*, 2003. 32(4): p. 352-7.

[35] Alterovitz, G., et al., Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry. *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology*, San Francisco, CA, USA, 2004.

[36] Anderson, A. and Z. Weng, *VRDD:* applying virtual reality visualization to protein docking and design. *Journal of Molecular Graphics & Modelling*, 1999. 17(3-4): p. 180-6.

[37] Fellenberg, M., et al. Integrative analysis of protein interaction data. in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 16-23 Aug. 2000*. 2000. La Jolla, CA, USA: AAAI Press.

[38] Han, K. and Y. Byun, Three-dimensional visualization of protein interaction networks. *Computers in Biology and Medicine*, 2004. 34(2): p. 127-39.

[39] D. Hirschberg S. Tryggvason M. Gustafsson M, Identification of endothelial proteins by MALDI-MS using a compact disc microfluidic system. *Protein Journal*, 2004. 23: p. 263-71.

[40] Mann, M.T.M., From genomics to proteomics. *Nature*, 2003. 422: p. 193-197.

[41] S.F. Altschul T.L. Madden A.A. Schaffer J. Zhang Z. Zhang W. Miller D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Research*, 1997. 17: p. 3389-402.

[42] B. Alberts A. Johnson J. Lewis M. Raff K. Roberts P. Walter, Molecular Biology of The Cell. 2002: *Garland Science*.

[43] T. D. Pollard W.C. Earnshaw, *Cell Biology*. 2002: W.B. Saunders Company.

[44] J.C. Venter M.D. Adams E.W. Myers P.W. Li et al, The sequence of the human genome. *Nature*, 2001. 291: p. 1304-51.

[45] US department of energy, The human genome project and beyond. 2003, US department of energy. p. 1-12.

[46] Leibler, Introduction to Proteomics: Tools for the New Biology. 2002, Totowa, NJ: Humana Press.

[47] D. Eisenberg E.M. Marcotte I. Xenarios T.O. Yeates, Protein function in the post-genomic era. *Nature*, 2000. 405: p. 823-26.

[48] BlackStock, W.P.W., M. P, Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 1999. 17: p. 121-127.

[49] Crick, F.H.C., Central dogma of molecular biology. *Nature*, 1970. 227: p. 561-563.

[50] K Machida, M.N., M Imaizumi, T Abe, Y Ohnishi, K Takagi, S Yoshii, M Hamaguchi, Tyrosine phosphorylation in lung cancer as a prognostic marker. *Cancer Detection and Prevention*, 1996. 5(20).

[51] Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. *Science*, 1998. 282(5396): p. 2012-8.

[52] S. P. Gygi Y. Rochon B. R. Franza R. Aebersold, Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*, 1999. 19: p. 1720-30.

[53] C. Hoog, M.M., Proteomics. *Annual review of Genomics and human Genetics*, 2004. 5: p. 267-93.

[54] Williams, V., Pathways of Innovation: a history of the first effective treatment for sickle cell anemia. *Perspect Biol Med*., 2004. 4(47): p. 552-63.

[55] Chapman, T., Automation on the move. *Nature,* 2003. 421: p. 661 - 66.

[56] Bachrach, C.A. and T. Charen, Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Med Inform* (Lond), 1978. 3(3): p. 237-54.

[57] Glover, J., Searching for the evidence using PubMed. *Med Ref Serv Q*, 2002. 21(4): p. 57-65.

[58] Bachmann, L.M., et al., Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc*, 2002. 9(6): p. 653-8.

[59] A.W. Dowsey M.J. Dunn G.Z. Yang, ProteomeGRID: towards a high-throughput proteomics pipeline through opportunistic cluster image computing for two-dimensional gel electrophoresis. *Proteomics*, 2004.

[60] M. Traini AA. Gooley K. Ou, Towards an automated approach for protein identification in proteome projects. *Electrophoresis*, 1998. 11: p. 1941-9.

[61] C.R. Mallet Z. Lu R. Fisk J.R. Mazzeo, Performance of an ultra-low elution-volume 96-well plate: drug discovery and development applications. *Rapid Communications in Mass Spectrometry*, 2003. 17: p. 163-70.

[62] www.thermo.com, *Thermo Electron*,. 2004.

[63] Choudum, S.A. and S. Sivagurunathan, Optimal fault-tolerant networks with a server. *Networks*, 2000. 35(2): p. 157-60.

[64] Mahgoub, I. and C.-J. Huang, A novel scheme to improve fault-tolerant capabilities of multistage interconnection networks. *Telecommunication Systems - Modeling, Analysis, Design and Management*, 1998. 10(1-2): p. 45-66.

[65] Yang, S.-C. and J.A. Silvester, Fault-tolerant multistage interconnection networks: performance/reliability tradeoffs. *Computer Systems Science and Engineering*, 1990. 5(4): p. 233-42.

[66] Arpinar, I.B., et al., Formalization of workflows and correctness issues in the presence of concurrency. *Distributed and Parallel Databases*, 1999. 7(2): p. 199-248.

[67] Ceroni, J.A. and S.Y. Nof, A workflow model based on parallelism for distributed organizations. *Journal of Intelligent Manufacturing*, 2002. 13(6): p. 439-61.

[68] Mahling, D.E., N. Craven, and W.B. Croft, From office automation to intelligent workflow systems. *IEEE Expert*, 1995. 10(3): p. 41-7.

[69] Rajakumar, S., V.P. Arunachalam, and V. Selladurai, Workflow balancing strategies in parallel machine scheduling. *International Journal of Advanced Manufacturing Technology*, 2004. 23(5-6): p. 366-74.

[70] H. Liu D Lin J.R. Yates 3rd, Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques*, 2002. 32: p. 898-902.

[71] P. Wickware P. Smaglik, Proteomics technology: Character references. 2001. 413: p. 869 - 875.

[72] Perkel, J.M., Technologies Vie for Dominance. *The Scientist*, 2003. 17.

[73] M. Yarmush A Jayaraman, Advances in Proteomic Technologies. *Annual review of Biomedical Engineering*, 2002. 4: p. 349-373.

[74] Mitra S Brukh R, *Sample Preparation Techniques in Analytical Chemistry*. 2003: John Wiley & Sons.

[75] Link AJ Eng J Schieltz DM Carmack E Mize GJ Morris DR Garvik BM Yates JR, Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 1999. 17: p. 676-82.

[76] Haynes P.A Yates J.R, Proteome profiling--pitfalls and progress. *Yeast*, 2000. 17: p. 81-87.

[77] Ruedi Aebersold, M.M., Mass spectrometry-based proteomics. *Nature*, 2003. 422: p. 198 - 207.

[78] G.A. Michaud M. Snyder, Proteomic Approaches for the Global Analysis of Proteins. *Biotechniques*, 2002. 33: p. 1308-16.

[79] John R Yates, Mass Spectrometry and the Age of the Proteome. *Journal of Mass Spectrometry*, 1998. 33: p. 1-19.

[80] M Merchant SR Weinberger, Recent advancements in surface enhanced lased desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*, 2000. 21: p. 1164-67.

[81] Petricoin, E.F., et al., Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 2002. 359(9306): p. 572-7.

[82] Dalmasso, G.R.E.A., SELDI ProteinChip Array Technology: Protein-Based predictive Medicine and Drug Discovery Applications. *Journal of Biomedicine and Biotechnology*, 2003. 4: p. 237-41.

[83] Zhang, L., Contribution of human a-defensin 1, 2 and 3 to the anti-HIV activity of CD8 antiviral factor. *Science*, 2002. 298: p. 995-1000.

[84] Marshall, A.G., *Accounts of Chemical Research*, 1996. 29: p. 308.

[85] Lahm H-W, L.H., Mass spectrometry: a tool for the identification of proteins separated by gels. 2000. 21: p. 2105-14.

[86] Mann M, H.R., Pandey A, Analysis of proteins and proteomes by mass spectrometry. *Annual. Rev. Biochem*, 2001. 70: p. 437-73.

[87] Kuster B, M.P., Andersen JS, Mann M, Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, 2001. 1: p. 641-50.

[88] Pappin DJ, H.P., Bleasby AJ, Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol*, 1993. 3: p. 327-32.

[89] Eng J, M.A., Yates JR, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*, 1994. 5: p. 976-89.

[90] Quadroni M, J.P., Proteomics and automation. *Electrophoresis*, 1999. 20: p. 664-77.

[91] JR Yates, Database searching using mass spectrometry data. *Electrophoresis*, 1998. 19: p. 893-900.

[92] Lill, J., Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom. Rev*, 2003. 22: p. 182-94.

[93] D.N. Chakravarti B. Chakravarti I. Moutsatsos, Informatic tools for proteome profiling. *Biotechniques*, 2002. Suppl 32: p. 4-15.

[94] D.N. Perkins D.J. Pappin D.M. Creasy J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 1999. 18: p. 3551-67.

[95] W Zhang B.T. Chait, ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry*, 2000. 72(11): p. 2482-89.

[96] J.R. Yates J.K. Eng A.L. McCormack D. Schieltz, Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database. *Analytical Chemistry*, 1995. 67: p. 1426-36.

[97] J.R. Yates J.K. Eng A.L. McCormack, Mining genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Analytical Chemistry*, 1995. 67: p. 3202-10.

[98] B.T. Hansen J.A. Jones D.E. Mason D.C. Liebler, SALSA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Analytical Chemistry*, 2001. 73: p. 1676-83.

[99] D.C. Liebler B.T. Hansen S.W. Davey L. Tiscareno D.E. Mason, Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Analytical Chemistry*, 2002. 74: p. 203-10.

[100] D.L. TaBB J.K. Eng J.R. Yates, Protein Identification by SEQUEST, in *Proteome Research: Mass Spectrometry*, P. James, Editor. 2001, Springer. p. 125-42.

[101] Jain, A.K. and J. Mao, Artificial neural networks*: A tutorial. IEEE Computer*, 1996. 29(3): p. 31-44.

[102] Cortes, C. and V. Vapnik, Support-vector networks. *Machine Learning*. 20(3): p. 273-97.

[103] Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning. 1989: Addison-Wesley.

[104] Hartigan, J. and M. Wong, Algorithm AS136: A k-means clustering algorithm. *Applied Statistics*, 1979. 28: p. 100-108.

[105] Joliffe, I., *Principal Component Analysis*. 1986, New York, NY: Springer-Verlag.

[106] Jain, A. and R. Dubes, Algorithms for Clustering Data. 1988, Englewood Cliffs, NJ.: Prentice-Hall.

[107] Cheeseman, P., Stutz, J., Bayesian Classification (Autoclass): Theory and Results, in Advances in Knowledge Discovery and Data Mining, G.P.-S. U. Fayyad, P. Smyth and R. Uthurusamy, Editor. 1996, MIT Press: Cambridge.

[108] Ramoni, M.F., P. Sebastiani, and I.S. Kohane, Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A*, 2002. 99(14): p. 9121-6.

[109] Troyanskaya, O.G., et al., *A* Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A*, 2003. 100(14): p. 8348-53.

[110] Gelman, A., et al., *Bayesian data analysis*. 1995, New York: Chapman & Hall.

[111] Heckerman, D., D. Geiger, and D.M. Chickering, Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 1995. 20(3): p. 197.

[112] Nigam, K., et al., Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000. 39(2-3): p. 103-34.

[113] Lee, Y. and C.K. Lee, Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 2003. 19(9): p. 1132-9.

[114] Wang, H., et al., Application of support vector machines to classification of blood cells. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, 2003. 20(3): p. 484-7.

[115] Alterovitz, G., et al., Machine Learning Techniques for Proteomic Classification and Marker Selection Using Sample Fractionation with SELDI-TOF MS. Presented at International Conference on Analysis of Genomic Data, Boston, MA, USA, 2004.

[116] Wu, B., et al., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 2003. 19(13): p. 1636-43.

[117] Byvatov, E. and G. Schneider, Support vector machine applications in bioinformatics. *Appl Bioinformatics*, 2003. 2(2): p. 67-77.

[118] Alterovitz, G., D.H. Staelin, and J.H. Philip, Temporal patient state characterization using Iterative Order and Noise (ION) estimatiun: applications to anesthesia patient monitoring. *J Clin Monit Comput*, 2002. 17(6): p. 351-9.

[119] Gay, S., et al., Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics*, 2002. 2(10): p. 1374-91.

[120] Conrads, T.P., et al., High-resolution serum proteomic features for ovarian cancer detection. *Endocr Relat Cancer*, 2004. 11(2): p. 163-78.

[121] Michalickova, K., et al., SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, 2002. 3(1): p. 32.

[122] Janeway, C.A., et al., *Immunobiology: The Immune System in Health and Disease*. 6th ed. 2004, London: Garland Science Publishing.

[123] Ross, S.M., *Simulation*. 2nd ed. 1997, Boston: Academic Press.

[124] Johnson, N.L., S. Kotz, and A.W. Kemp, Univariate Discrete Distributions. 2nd ed. *Wiley Series in Probability and Statistics*. 1993, NY: Wiley-Interscience.

[125] Miaou, S.P. and H. Lum, Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention*, 1993. 25(6): p. 689-709.

[126] Prince, J.T., et al., The need for a public proteomics repository. *Nat Biotechnol*, 2004. 22(4): p. 471-2.

[127] Spitz, M.R., et al., Association between malignancies of the upper aerodigestive tract and uterine cervix. *Head Neck*, 1992. 14(5): p. 347-51.

[128] Friedl, F., et al., Studies on a new human cell line (SiHa) derived from carcinoma of uterus. I. Its establishment and morphology. *Proc Soc Exp Biol Med*, 1970. 135(2): p. 543-5.

[129] www.chip.org/proteomics/raspap.

[130] Kohavi, R., The Power of Decision Tables. *Proc European Conference on Machine Learning*, 1995.

[131] Frank, E., et al., Data mining in bioinformatics using Weka. *Bioinformatics*, 2004. 20(15): p. 2479-81.

[132] Wheeler, D.L., et al., Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, 2003. 31(1): p. 28-33.

[133] Sasaki, H., et al., Cten mRNA expression is correlated with tumor progression in thymoma. *Tumour Biol*, 2003. 24(5): p. 271-4.

[134] Al-Maghrebi, M.A., F. Al-Mulla, and L.T. Benov, Glycolaldehyde induces apoptosis in a human breast cancer cell line. *Arch Biochem Biophys*, 2003. 417(1): p. 123-7.

[135] Jalbout, M., et al., Polymorphism of the stress protein HSP70-2 gene is associated with the susceptibility to the nasopharyngeal carcinoma. *Cancer Lett*, 2003. 193(1): p. 75-81.