PROTEOMICS

GIL ALTEROVITZ
Massachusetts Institute of Technology
Cambridge, Massachusetts
EHSAN AFKHAMI
Boston University
Boston, Massachusetts
JOSEPH BARILLARI
Harvard University
Cambridge, Massachusetts
MARCO RAMONI
Harvard Medical School and Harvard Partners Center for Genetics and Genomics
Boston, Massachusetts

1. INTRODUCTION

Research in proteomics involves studying the structure, expression, localization, interactions, and cellular roles of all proteins within a particular organism or subcomponent thereof (1). Researchers coined the term "proteomics" in the early 1990s (2) to describe a new approach to studying proteins, focusing on high-throughput analyses and on breadth rather than depth. Proteomics researchers aim to use experimental techniques that trade accuracy for volume in order to build up a complete picture of the function of large groups of proteins. Key research in the field focuses on the development of new high-throughput techniques and the computational machinery needed to analyze the data those techniques produce.

Proteomics has the potential to dramatically impact medicine. Scientists at the National Institutes of Health made headlines when they announced in 2002 that they could diagnose ovarian cancer using mass spectrometry-based proteomics (1). The pharmaceutical industry is also heavily involved in proteomic research. As most drugs target and inhibit the functions of specific proteins, drug discovery benefits greatly from proteomic assays that permit the identification or quantification of many proteins simultaneously.

As of this writing (2005), proteomics is growing the way genomics grew in the 1990s, when a series of sequencing projects created an ocean of genome sequences for researchers to analyze. In addition, the number of those genetic sequences in Entrez (a database of molecular biology-related information) is starting to saturate, whereas the number of proteins being cataloged in Entrez is still growing exponentially each year. This growth suggests that increasingly advanced techniques will be needed to deal with ever-larger proteomic datasets.

Although much of the engineering and statistical methodology developed for functional genomics (3) can be recycled for use in proteomics, the field has no shortage of interdisciplinary problems amenable to attack by researches ranging from electrical engineers to biophysicists. A few of the open problems include the fabrication of effective, accurate protein arrays (instruments to measure protein expression; see the section of the same name below) (4), design and construction of robots to automate repetitive tasks (5), and novel machine learning techniques for data analysis (6).

2. PROTEOMICS: MOLECULAR AND CELLULAR BIOLOGY FOUNDATIONS

This section summarizes some core molecular and cellular biology concepts that underlie the study of proteomics (7). Proteins are the biochemical machines responsible for life. Proteins read, copy, and organize the genetic code stored in DNA, digest nutrients, attack pathogens, and direct growth. Protein-based signals enable cells in a multicellular organism to communicate; structural proteins hold that organism together. Many open research problems in modern biology and medicine are, fundamentally, questions about the functions of proteins.

A protein is a chain of linked amino acids, the precise ordering of which determines its structure and function. Amino acids are biomolecules with four invariant components: a central carbon atom, to which the other components are bound; a hydrogen atom; an amino group (NH_2); and a carboxyl group (COOH). A variable component, called the R group, determines the type of each amino acid (the hydrogen atom below the carbon is the R group in Fig. 1). There are 20 standard R groups (for the 20 amino acids) and several additional variants. In a protein, the amino group of each amino acid is linked to the carboxyl group of the next amino acid, forming a chain. Some proteins have a few tens of amino acids, others, such as *huntigtin*, responsible for Huntington's disease, are composed of thousands.

Proteins can be categorized by shared functions, structures, or subsequences. Common functions include handling metabolic chores, providing structural support, and participating in signaling pathways, to name a few. The common structures α-helix (a twisted chain of amino acids) and β-sheet (a plane made of adjacent chains of amino acids) are often observed as components of larger structures. Common (well-conserved over the course of evolution) subsequences are regularly used to track evolution of organisms. Conserved sequences and the structures to which they give rise can be thought of as being made of modular units (called "motifs" or "domains") that confer specific properties and functions. Some motifs are well preserved across millions of years of evolution, through many different organisms. Many examples of preserved functions may be found in Gene Ontology, a controlled vocabulary of common functions (*http://www.geneontology.org*). Branden and Tooze's *Introduction to Protein*

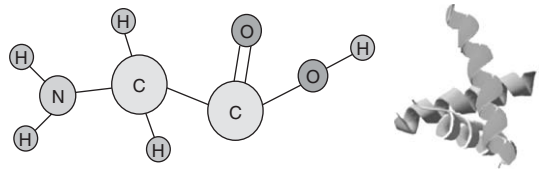

Figure 1. Examples of amino acid and protein structure.

Structure (8) contains many examples of highly conserved structures.

Approximately 40% of the human genome encodes proteins with no known function (9). Assigning functions to these proteins and their interactions is one challenge of proteomics.

Proteins are created in a two-step process. First, DNA is *transcribed* into RNA. Then, the RNA is *translated* into protein. Both transcription and translation are copying processes; neither involves changing the DNA or RNA template for the copying. This process, in which information flows from DNA to RNA to protein, is called the "central dogma" of molecular biology. Any time after it is translated, a protein may be altered by a variety of *posttranslational modifications*. The study of these modifications is another major component of proteomics.

Nearly every cell in an organism contains a copy of that organism's complete genome, but each cell expresses (synthesizes proteins from) only a subset of that genome. Genes that encode proteins essential to basic cellular functions are expressed in nearly all cells, whereas those with highly specialized functions are expressed only in certain cell types. (Some proteins found in liver cells would never be expressed in the brain, for instance.) These subsets of expressed proteins may be called proteomes. Although every organism has one genome, a multicellular organism may have many proteomes.

Analysis of the most recent version of the human genome suggests that humans have between 20,000 and 25,000 genes—only slightly larger than the approximately 19,000 genes in the genome of the worm *Caenorhabditis elegans* (10). This suggests that the vast difference in complexity between humans and worms cannot be explained merely by the fact that humans have more genes. Proteomics may provide one means of explaining that gap, by showing how the interactions between proteins give rise to human complexity.

3. PROTEOMICS: METHODS AND TECHNOLOGIES

3.1. Overview of Proteomic Technologies

The workhorse of proteomics is the mass spectrometer, an instrument used to count and measure the mass-to-charge ratios of ions (charged particles). The following sections discuss how proteomics researchers inject proteins into a mass spectrometer, how they are measured inside the instrument, and how one can use the gigabytes of data produced by the instrument to identify, sequence, and quantify proteins.

Mass spectrometry (MS) is not the only analytical method in proteomics. After the sections on MS is a section on protein array analysis, which is another technology used to identify and quantify (but not sequence) many proteins simultaneously. Further sections discuss protein databases and the discovery of protein–protein interactions using laboratory techniques such as yeast two-hybrid analysis and computational techniques such as data mining of the literature.

3.2. Approaches to Mass Spectrometry-Based Proteomics

Research in mass spectrometry has grown rapidly in recent years. The field of mass spectrometry in general has grown over 2.5 times over the past decade in terms of PubMed-related publications measured as discussed below. This compares with a one-third 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 classic analytical chemistry-affiliated molecular studies) such as proteome mining, post-translational modifications (PTMs), and protein–protein interactions. The immense amounts of data generated by mass spectrometry-based proteomics have paved the way for systematic identification of proteomes and intracellular dynamics. Mass spectrometry is also easily adaptable to high-throughput formats, a fact that has made it the method of choice for protein identification and characterization (11). Although an exhaustive review is beyond the scope of this article, the following will give an overview of the relevant technology and biomedical applications within the context of this section.

There are three main components in any mass spectrometer: the ion source, the mass analyzer, and the detector. The source produces ions from the biological sample, the mass analyzer resolves the ions [in a mass-to-charge (m/z) ratio-dependent manner], and the detector detects the ions resolved by the mass analyzer. From an ion's point of view, mass spectrometry converts a sample into ions, groups those ions by mass-to-charge ratio, and measures the intensity of each collection of ions with a common m/z ratio.

The most straightforward use of mass spectrometry in proteomics would be to ionize a mixture of proteins, spray it into a mass spectrometer, and use the mass-to-charge ratios to identify and quantify every protein in it. This approach, called "top-down" proteomics, is not without its proponents, as modern instruments are becoming increasingly accurate with the large masses involved in top-down experiments (12,13). If every protein had a unique mass and mass spectrometers were absolutely accurate, then one would need no other methods. However, in contemporary mass spectrometers, measurement accuracy decreases as the absolute mass increases, making accurate identification of large proteins difficult. Many different proteins may have masses within the margin of error for these measurements. PTMs, discussed above, further muddy the water—many PTMs change the mass of a protein but do not change its sequence. An active area of research involves looking at the statistical issues involved in top-down protein identification.

An alternative approach is "bottom-up" or "shotgun" proteomics, in which proteins are chopped into peptides (short sequences of amino acids) before identification, a process called "digestion." Bottom-up proteomics has three major advantages over the top-down approach. First, as mass spectrometers are more accurate for smaller masses, they are better at resolving small peptides rather than large proteins. Second, the bottom-up approach also greatly reduces the chance that PTMs will trip up the identification process: If enough peptides are unmodified,

the protein can be identified, regardless of how many modifications were made to the other peptides. Finally, in tandem mass spectrometry (in which select ions are broken into fragment ions and the fragments are sent for another round of mass spectrometry), the bottom-up approach yields easier-to-analyze fragment spectra because peptides have fewer components to break apart than do intact proteins.

In the bottom-up approach, the peptides are sprayed into the mass spectrometer and their *m/z* ratios are measured. Trypsin, the protease most commonly used to digest protein samples into peptides, cleaves proteins at very predictable amino acid locations. If one knows the genome sequence of the organism that provided the protein sample (which is the case for most model organisms used in biological research), one can calculate the mass of all possible fragments from all of the proteins in the organism. In the process of peptide mass fingerprinting (PMF), the unknown protein of interest is cut into peptides by an enzyme such as trypsin. The absolute mass of these peptides is measured with a mass spectrometer. Using software, these masses are then compared with the theoretical masses of peptides coming from that organism. This process demands high sensitivity, resolution, and accuracy (14). Sensitivity is required to measure masses on the order of femtomole (10^{-15}) quantities with high resolution to distinguish between ions of similar *m/z* values. Although some peptide sequences of approximately six or more amino acids in length would have unique masses within the proteome of an organism, using additional peptide fragments can improve confidence in the identification (15); in other words, a protein from which several peptides were identified is more likely to be present than one that had only one successful "hit."

Although mass spectrometry is a sensitive method for identifying proteins, it is more difficult to use mass spectrometry to accurately quantify proteins (16). 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 experimental samples) while leaving the other unchanged (e.g., the control samples). This modification changes the molecular weight of the isotope-based samples relative to controls, but not the mass spectrometer's behavior in terms of the peak intensities. Quantitative differences in proteins are then determined directly as the difference in peak area between the two peptides in the mixed samples (i.e., control and cancer) (17). A well-illustrated overview of the techniques of mass-spectrometry-based proteomics can be found in the reference section of this article (18).

3.3. Analytical Polypeptide Separation (2-D SDS-PAGE and HPLC)

Most biological samples (serum, blood, urine, and cell lysates, to name a few) cannot be sprayed directly into a mass spectrometer. First, biological samples often contain a considerable amount of non-protein material, which must be removed. Centrifugation is a common means of removing the largest non-protein components (for instance, the cell debris left over in a cell lysate). Many samples need further processing to remove salts and other small-molecule contaminants.

The resulting pure-protein mixtures are also often too complex for direct analysis with a mass spectrometer; if they were analyzed all at once, the sheer quantity of proteins would overload the detector. Proteomics researchers avoid this problem by separating proteins in advance according to their physical or chemical properties.

Popular protein separation methods include two-dimensional (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 mass spectrometry (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 (although 2-D SDS-PAGE is still prominent today (17)).

In 2-D SDS-PAGE, proteins are loaded onto a gel and subjected to an electric field. The chemical properties of the gel prompt the proteins to separate in one dimension by their isoelectric point (i.e., the pH where protein has zero net charge) and in the other dimension by their 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. A practical application of this method would be to compare samples from differing cellular states (diseased and normal). This comparison can give scientists insight as to which proteins differentiate the two states and should be further investigated. SDS-PAGE has some major shortcomings. Generally, if a protein mixture is to be characterized in an SDS gel by MS, it requires some partial purification to reduce complexity before analysis. Despite significant technical improvements, protein separations patterns are often not reproducible. Also, SDS gels perform poorly in detecting low abundance proteins.

Integrated systems for performing 2-D SDS-PAGE are entering the marketplace. Contemporary systems include facilities for robotic sample preparation, 2-D gel electrophoresis, gel extraction via precision robots, ionization labeling, and mass spectrometry peptide fragments analysis. In these systems, data generated from all instruments are presented with using a graphical user interface. These systems are useful for high-throughput analysis, contributing to significant increases in processing power (19). There are major shortcomings, however, in such systems. An example is the homogenous treatment of samples with no feedback control mechanism. For instance, a laboratory technician doing a gel protein digestion would match the amount of protease (an enzyme used to cleave the protein into peptides) used to digest a spot on

the gel to the amount of protein in the spot by observing the spot's intensity directly with the naked eye. Intelligent systems capable of such adjustments have not yet reached the market (19).

In bottom-up proteomics, 2-D SDS-PAGE is commonly used before protein digestion. By contrast, HPLC is commonly used to separate the peptides resulting from a digestion (in which proteins are chopped into smaller pieces with a protease such as trypsin). HPLC involves pumping the peptides through a chromatography system that gradually releases them over a time interval (typically in the range of an hour) depending on their physical or chemical characteristics. Some characteristics used include

- 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 systems, also known as tandem liquid chromatography (LC/LC) systems, pump a sample through two or more steps of LC to separate the peptides based on multiple attributes. Multidimensional LC coupled with tandem mass spectrometry (LC-LC-MS/MS) is used in the analysis of very complex mixtures of peptides, in which the additional LC step reduces the number of peptides entering the mass spectrometry at the same time. This method is commonly known by the acronym Multi-Dimensional Protein Identification Technique, or MudPIT (20).

LC is not without its challenges however. A major limitation of HPLC is that one cannot generally achieve the chromatographic resolution provided by some other forms of chromatography such as gas chromatography. Also, HPLC is not readily interfaced with a mass spectrometer because the liquid phase presents problems with the high vacuum required for mass spectral analysis. However, progress has been made in this area and LC-MS has become a vital tool in many proteomic laboratories.

3.4. Ionization Methods

Three prominent mass spectrometry ionization methods used in proteomics are Electrospray Ionization (ESI), Matrix Assisted Laser Desorption and Ionization (MALDI), and Surface-Enhanced Laser Desorption and 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 sprayed into the mass analyzer. The mist is often the output of an 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 (21). MALDI mass spectrometry is normally used to analyze relatively simple peptide mixtures, whereas integrated HPLC 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 (22). Each peptide tends to (generally) pick up a single proton, creating a positive ion. This is significant because 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). 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 select 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 (23) or in creating diagnostically useful proteomics profiles.

SELDI has become increasingly popular since a study from Liotta et al. was first published in *Lancet* (24,25) involving diagnosis of ovarian cancer without actually identifying any proteins. As shown in Fig. 2, 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 1990s. The subset of MALDI/SELDI papers affiliated with proteomics has exhibited even faster growth.

As alluded to earlier, mass spectrometry is also a clinical tool and has been used in numerous disease studies (25–27). In an HIV study (28), 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, recent studies (25,29) predicted 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 based on the overall pattern of protein expression. In the case of ovarian cancer, the importance of early diagnosis is apparent in the high 5-year survival rate (95%) of patients with cancer limited to the ovary compared with the low 35–40% 5-year survival rate for late-stage patients (25). SELDI has also been used in diagnosis of neurological diseases such as Alzheimer's disease, Parkinson's disease, multiple sclerosis, schizophrenia, and many others (27).

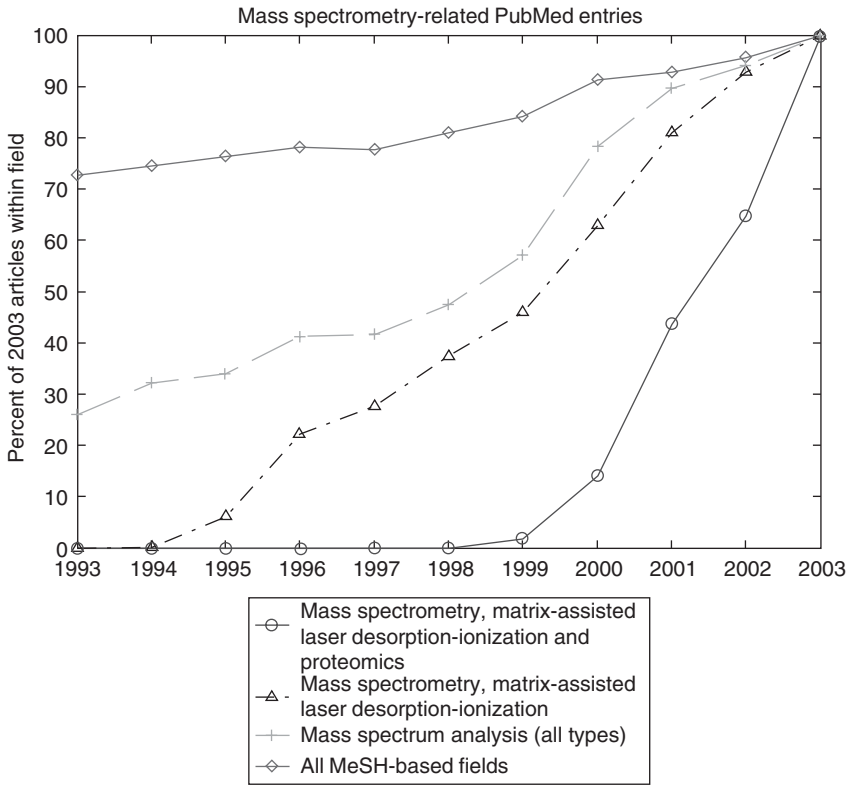

Figure 2. Mass spectrometry is growing at a much faster rate in terms of papers compared with the general PubMed database.

3.5. Mass Analyzers

Three basic types of mass analyzers are currently used in proteomics research: the ion trap, time-of-flight (TOF), quadrupole time-of-flight (Q-TOF), and Fourier transform (FT-MS) ion analyzers. Each is different in design and performance, and each has its advantages. They can be used alone or arranged in tandem to take advantage of the unique strengths of each in tandem mass spectrometry (11).

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 mass spectrometry or tandem mass spectrometry (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 (11). The FT-MS is similar to an ion trap, but it employs a magnetic field for detecting ions in the trap (30). FT-MS instruments have high resolution and are excellent for measuring low abundance proteins as well as complex peptide mixtures. This is because the superconducting magnet has a stability of resolving few peptides among a billion. Quadrupole and TOF instruments can at best deliver a 1 per 10,000 resolution. However, current models of FT-MS (owing, among other factors, to their need for a cryogenically cooled superconducting magnet) are extremely expensive and operationally complex. This, coupled with their low-peptide-fragmentation efficiency, has limited the use of FT-MS in proteomics research (22). TOF analyzers (Fig. 3) measure the time the gas-phase ions take to travel from the ionization source to the detector, which is used to calculate to the *m/z* ratio (31). TOF analyzers are not well suited for MS/MS (see below) and have the disadvantage of being dependent on sample quality for successful peptide identification (11). 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), whereas 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:

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \tag{1}$$

where:

t = time of flight (μs)
m = mass (Da)
z = charge (C)
U = voltage (e.g., 20,000 V)
a, b, c = model constants (e.g., $a = 0.272$, $b = 0$, $t_0 = 0.0038$).

An overview of tandem MS (MS/MS) is shown in Fig. 4. First, peptide ions generated from an ESI source are separated based on their *m/z* ratios. In the second round, a single *m/z* is chosen and is subject to collision-induced dissociation (CID)—the ions of that *m/z* are bombarded with a charged gas, which causes them to fragment (32).

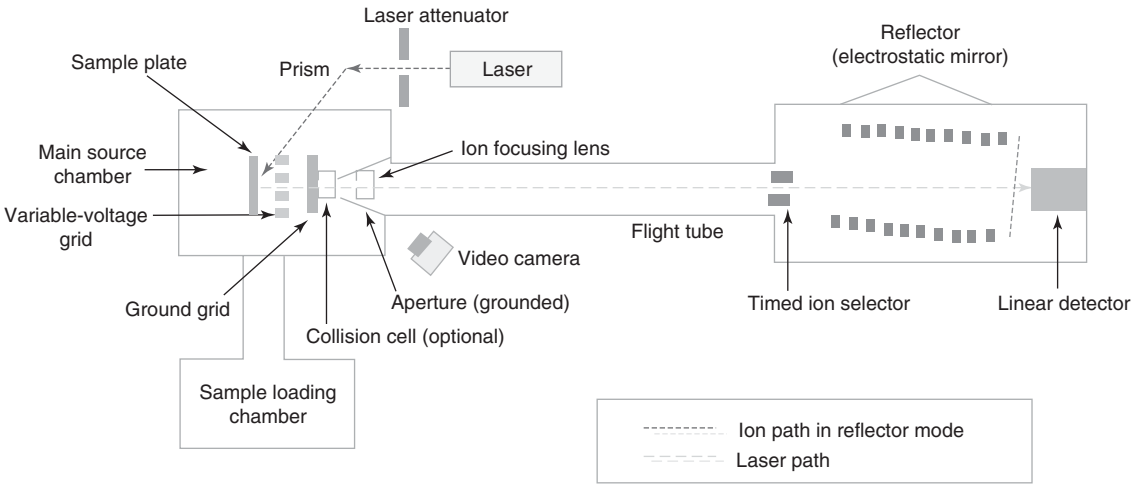

Figure 3. SELDI-TOF Mass Spectrometry schematic.

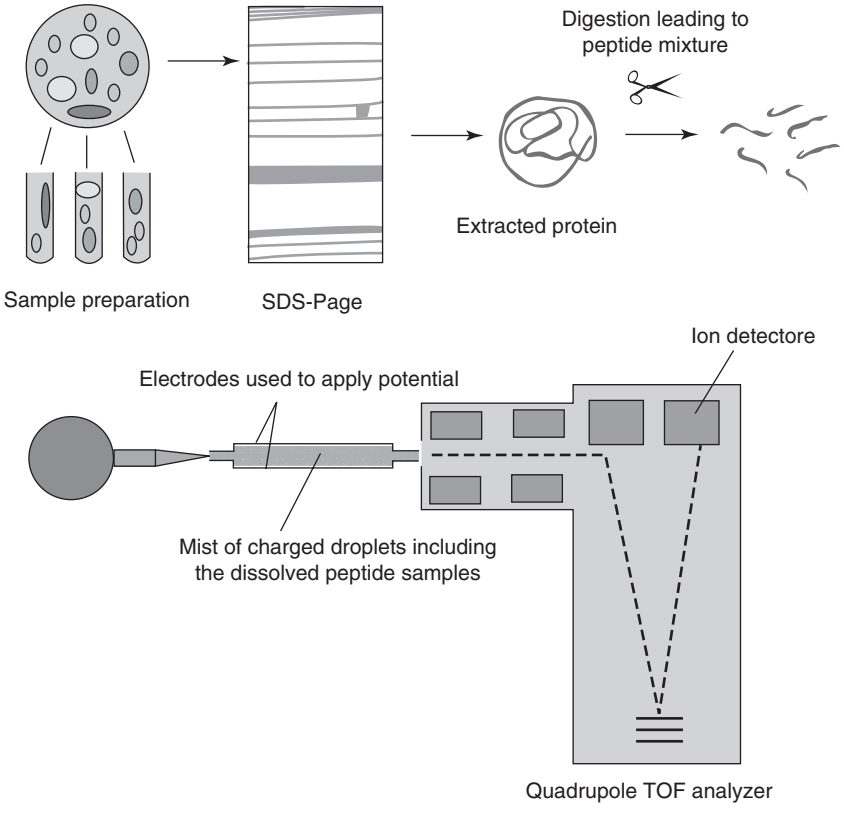

Figure 4. Steps involved in tandem mass spectrometry.

The fragments are then scanned with the mass spectrometer. The resultant tandem spectra of amino acid composition can be searched against protein databases to identify the protein (33). Matches from at least two peptides derived from the same protein are typically required to positively identify a protein (34), with each additional match adding confidence to the identification. Computational analysis of MS/MS spectra can also provide information about the nature and location of peptide modifications (35). 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, is the most widely used tool for searching protein databases (36). Sequest, which is discussed further in the next section, is well suited for high-throughput proteomics as it automatically extracts and searches the MS/MS data against a protein database (37). Other database-search tools include the newer Mascot, published by Matrix Science and available for use on the

Web, and the open-source OMSSA, published by the National Center for Biotechnology Information.

3.6. Identification of Proteins from MS and MS/MS Data

With the output of an MS or MS/MS run in hand, one may employ a variety of methods to try to identify the proteins in the sample injected into the mass spectrometry instrument. Popular approaches include peptide mass fingerprinting (searching databases for the masses of peptides), MS/MS database searching (searching databases of theoretical MS/MS spectra, as does Sequest), and sequence tag searching [partial sequences ("tags") are derived from MS/MS spectra and used to query sequence databases].

In a peptide mass fingerprinting approach, a protease is applied *in silico* (in other words, virtually) to all entries in a protein database (e.g., Swiss-Prot, OWL, or NCBInr) to yield a list of peptides with corresponding theoretical masses. Matches are made between observed peptide masses obtained from MS and the theoretical masses from the database. If several of these peptides uniquely match the same protein, then the unknown sample protein can be identified. If there are multiple proteins in the sample (as there often are), 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 (15). Some probability-based scoring systems have emerged (38); one such algorithm is ProFound (39).

De novo sequencing involves measuring the distances between peaks in the MS/MS spectrum of a fragmented peptide, looking for distances that correspond to the mass of a single amino acid (most amino acids have distinctly sized masses; see Table 1), and chaining these together to form a partial sequence. GutenTag is a well-known program that implements this approach; experienced scientists can sequence spectra by hand (albeit more slowly) (40).

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. As 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. 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, the peptides of a specific *m/z* are selected and subject to CID, which breaks them into fragments. The fragmentation process primarily gives rise to cleavage products that break along peptide bonds. Because of this simplicity in fragmentation, it is possible to use the observed fragment masses to match with a database of predicted masses for one of many given peptide sequences. As an example, the peptide GVAGNEGAL might be fragmented into GVAG and NEGAL. 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. The spectra of the fragments (in which the fragments become peaks) can then be analyzed to obtain the sequence by looking for the aforementioned gaps between peaks that are the same size as an amino acid and using them to reconstruct a partial sequence tag. The *de novo* method is usually followed by a search of an *in silico* digested protein database, similar to PMF, to identify the protein the peptide originated from.

A third approach to determining the sequences of peptides is to use MS/MS data to search databases of synthetic peptide digests. Sequest (41) and Mascot (*http://www.matrixscience.com*) are two widely used programs that employ this approach. Sequest's approach generates identifications using two pieces of information: the *m/z* ratio of the peptide before fragmentation (obtained from the first mass spectrometry step) and the MS/MS spectrum. Sequest looks up the *m/z* value of each peptide

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

being analyzed in a master list of peptides generated from a computationally digested protein database, as in peptide mass fingerprinting. Unlike PMF, Sequest's approach determines the peptide's identity by comparing the theoretical MS/MS spectrum of each peptide in the list with the observed MS/MS spectrum. (Sequest creates the theoretical MS/MS spectra from these peptides with a model of how peptides fragment in the CID process.) Sequest assigns a cross-correlation score (XCorr) to each theoretical peptide; the XCorr is used to select the best match.

3.7. Protein Arrays

Mass spectrometry is not the only high-throughput means of identifying the proteins present in a sample: Protein arrays provide an alternative approach. A prototypical protein array consists of a set of probes bound to a surface. Protein is applied to the surface and then washed away, so that the proteins that did not stick to any probes are removed. A variety of probes might be used—antibodies (to test for the presence or abundance of proteins), other proteins (to assess interactions), or drugs, other small molecules, DNA, RNA, and substrates for enzymes (to test binding). A single array might contain thousands of probes.

Compared with traditional methods of surveying protein binding and interactions, protein arrays are highly parallel and are often miniaturized. Their advantages include speed, high sensitivity, economical reagent usage, and abundance of data generated per experiment.

Array technology was first developed as a tool for high-throughput gene expression analysis. By combining small sample volumes and the ability to generate massive amounts of data in a single experiment, gene expression arrays have vastly accelerated the search for functional effects of single nucleotide polymorphisms (SNSPs) and modified gene expression in normal and diseased states. Much interesting science has come from the study of gene expression arrays. However, many array studies operate under the assumption that changes in mRNA levels ultimately correlate to changes in encoded protein levels. This assumption is in many cases incorrect (42)—gene expression analysis is no substitute for protein expression analysis. (As biochemical changes in the cell are generally correlated with the actions of protein, scientists tend to be more interested in the latter than the former.) Gene expression arrays also cannot provide information on protein PTMs, something that a properly designed protein array could do.

Protein arrays are typically built by immobilizing proteins (or other probes, such as small molecules) on surfaces such as glass, membranes, micro-liter wells, mass spectrometer plates, and beads (or other particles). A schematic for differential protein expression profiling with a fluorescence detection system is shown in Fig. 5. The surface chemistry of the array is designed to immobilize the surface molecules. The target proteins are exposed to binding molecules on the array. A detection system is then used to indicate the abundance of the target proteins. (One method might involve fluorescently labeling the target proteins and scanning the array for fluorescence after washing away an unbound protein.) Depending on the experimental design, some software (and even some hardware) in a protein array experiment can be adapted from machinery for DNA arrays.

Protein analyte-antibody binding may be detected directly or via a secondary antibody in a sandwich assay (Fig. 6). Direct labeling can be used for comparing distinct samples using different fluorophores. The differences in the target protein concentrations (within each capture spot) can be then detected via wavelength fluorescence analysis (43). (This is similar to a common method with DNA microarrays: The control and experimental sample are labeled with different fluorescent colors. Both are applied to the array, and the excess is washed off; the relative color of each probe is assessed to see which sample bound more strongly.) Sandwich immunoassays are the method of choice (providing high specificity/sensitivity) for low-abundance proteins (femtomolar range (44)) when antibodies for the protein are available (Fig. 7). This method can also be used for the detection of protein

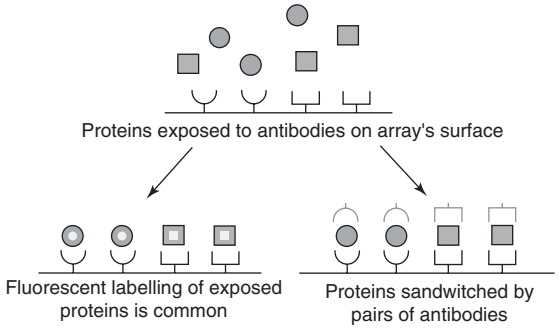

Figure 6. Capturing proteins.

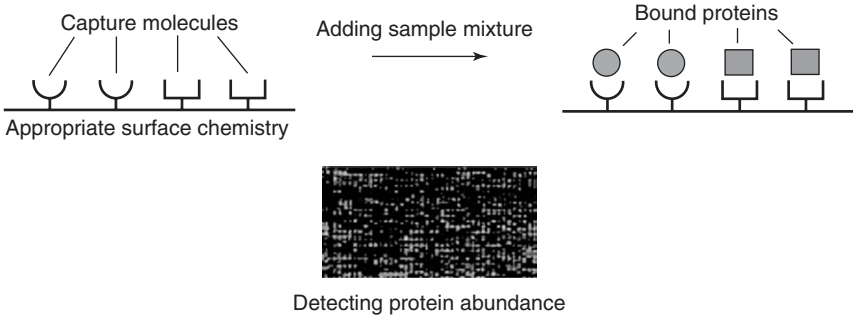

Figure 5. Protein array detection system.

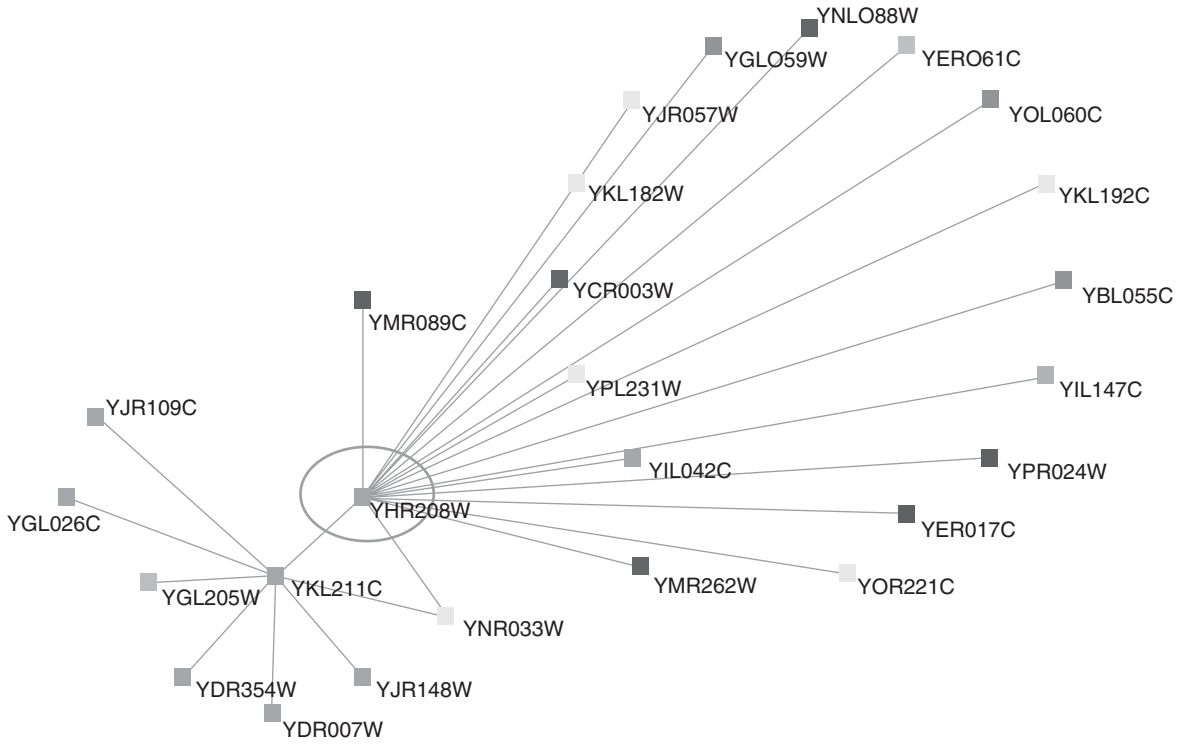

Figure 7. Random, scale-free, and hierarchical network architectures.

modifications. Cross-reactivity is an important issue for this technology. Although antibodies are conceptualized as being highly specific, unpredictable cross-reactions are possible. Thus, the usefulness of individual reagents depends on the relative level of cross-reaction and specific reaction. The use of sandwich assays, in which antibody pairs are used to bind and detect proteins, is one solution to this issue. This adds specificity because it is unlikely that both members of the sandwich will exhibit the same cross-reactivity. In summary, the factors required from such detection methods involve optimal sensitivity and specificity, with low background noise to give a high signal-to-noise ratio.

3.8. Yeast Two-Hybrid

Yeast two-hybrid (Y2H) is a molecular genetic technique that is commonly used for high-throughput mapping of potential protein–protein interactions. In its simplest form, the transcription of a reporter gene (e.g., β-galactosidase) is to signal that a (prey) protein has attached to a second, bait, protein. To accomplish this, a multidomain transcriptional activator of this reporter gene (e.g., Gal4) is used. Hybrid proteins are produced in which the bait is attached to one of these domains (i.e., DNA-binding domain), whereas the other is attached to the second domain (i.e., activating domain). If the bait and prey proteins bind, then the transcriptional activator can function, and this results in transcription of the reporter protein (which can then be measured). In this way, multiple bait proteins can be screened against a large array of prey proteins to find out which ones bind.

There are many related engineering issues in Y2 H. The technology allows for high-throughput instrumentation design and analysis. Improving the quality of the interaction measurement is another area of research. There are currently many false-positives and false-negatives. In fact, studies have estimated 50–90% are false-positives (45). "Sticky proteins" may also bind to many proteins without being biologically relevant. Technology limitations may lead to the misfolding of proteins, which then fail to interact (46).

3.9. Proteomic Databases

Extensive information on proteins gathered both from proteomics experiments and from experiments in the pre-proteomics era is available from public online databases. One can roughly categorize the major databases of interest to proteomics researchers as those containing sequence data, structure data, interaction data, mass spectrometry data, and the integration of the aforementioned data.

This section introduces the general content of each database type and refers to the most popular databases of each category. It should be noted that there are few globally accepted standards for database structure and implementation. Would-be database integrators often run into the perennial problem with biological databases: extensive redundancy and the lack of a common naming system to help match records and remove redundancy.

3.9.1. Protein Sequence Databases. At their core, most protein sequence databases contain the amino acid se-

quence of identified proteins. Some databases also include identification tags and references to a related journal article. Entrez and Swiss-Prot are among the most popular sequence databases.

Entrez (47) is a molecular sequence retrieval system developed at the National Center for Biotechnology Information (NCBI). Entrez Protein, a protein sequence database, is just a small subunit of the Entrez system. Entrez also provides access to biomedical literature, nucleotide sequence databases, three-dimensional (3-D) molecular structures, complete genome assemblies, OMIM (Online Mendelian Inheritance in Man, a database of genetic diseases), and many other resources.

Swiss-Prot (48), another popular protein sequence database, was established in 1986 through collaborative efforts of the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). The Swiss-Prot system relies on the translations of DNA sequences from the EMBL Nucleotide Sequence Database. EMBL is a comprehensive database of DNA and RNA sequences collected from the scientific literature, patent applications, and submissions directly from researchers/sequencing groups. TrEMBL is a computer-annotated supplement of Swiss-Prot that contains translations of EMBL nucleotide sequence entries (before being integrated into Swiss-Prot). Swiss-Prot is known for a minimal level of redundancy and a high level of integration with other databases.

3.9.2. Protein Structure Databases. Protein structure databases contain 3-D structural (e.g., secondary and/or tertiary) information. One of the best known is the Protein Data Bank (PDB) (49), an international repository of experimentally determined 3-D structures of biological macromolecules. The repository includes atomic coordinates (typically determined using X-ray crystallography, a highly accurate means of determining protein structure), bibliographic citations, secondary structure information, crystallographic structure, and nuclear magnetic resonance (NMR) experimental data.

3.9.3. Protein Interaction. Many databases collect lists of protein–protein interactions. The Database of Interacting Proteins (DIP) (50) is a database of pairs that are known to interact (e.g., two amino acid chains that bind to each other). DIP contains the name and the PIR/SWISSPROT/NCBI/EMBL unique identifier for each protein and any available information about the interaction. This may include the region involved in the interaction, the dissociation constant, and the experimental methods used to study the interaction.

BIND (51) is an another major interaction database. It has three classifications for molecular associations: molecules that associate with each other to form interactions, molecular complexes, and pathways. Complexes are functional combinations of two or more molecules, capable of performing a specific function. Pathways are a sequence of temporal events (interactions) that occur within cells. In BIND, complexes and pathways are represented by molecular complex objects and pathway records, respectively, both of which are formed by linkage of two or more interaction records.

The KEGG database (52) integrates data on molecular interaction networks in biological processes as well as chemical compounds and reactions. Metacyc/Ecocyc (53) is another database that collates metabolic and other regulatory pathway information.

A recent new development in proteomics databases is the Proteomics Standards Initiative (PSI) standard (54). This initiative aims to define community standards for data representation in proteomics. PSI is taking steps to standardize mass spectrometry and protein–protein interaction data. The PSI–MI (molecular interactions) format is a data exchange format for protein–protein interactions. Although that initiative seeks to standardize the structure of databases, the actual content is left ambiguous. Also, data in its fields can vary somewhat across databases: In databases supporting the PSI–MI format, the proteins may be referenced by different identifiers ranging from Uniprot identifiers, NCBI GI numbers, Ensembl identifiers, and the International Protein Index (IPI). In addition, virtually no database contains all fields in the PSI–MI specification.

Still more interaction databases have been drawn from literature mining. In a literature-mining approach, text processing software is applied to a large database of biomedical literature (the NCBI's PubMed abstracts, for instance) to glean protein–protein interactions described in the text. One recent approach combined text-mining with some of the experimentally derived databases described above (55).

3.9.4. Mass Spectrometry Databases. There are a few nascent public mass spectrometry databases at this time. The Open Proteomic Database (OPD) (56) and Peptide Atlas Repository are two such examples. The OPD, at the University of Texas-Austin, is roughly a collection of 1,200,000 spectra representing experiments from four different organisms. The Peptide Atlas Repository (Institute for System Biology) contains the same type of data, with additional quantitative filtering methods applied to the received data.

3.9.5. Integration Databases. SeqHound (57) and AliasServer (58) are well-known examples of integration databases, which combine data from multiple sources. SeqHound combines sequence and structural information with additional annotation data on the biomolecules in its catalog. AliasServer provides a cross-reference service that links the many different identifiers used by different databases to refer to the same biomolecules. Both SeqHound and AliasServer provide an application programmer interface (API) to aid the creation of computer programs that access the databases over the Internet.

4. MODELING PROTEIN NETWORKS AND INTERACTIONS

Mass spectrometry, protein arrays (59), and the yeast two-hybrid technique (60) can produce reams of raw protein–

protein interaction data. Making sense of these data is a major computational challenge in proteomics.

A natural representation for a collection of protein–protein interactions is a graph. By way of review, a graph G consists of a nonempty set of vertices V and a set of edges E that potentially link vertices together. $G=(V,E)$ where $E=\{(u,v)|\ u,v\in V\}$. A graph may be directed or undirected. A directed graph is one in which each edge has a direction (in other words, an edge from a to b is distinct from an edge from b to a). Conversely, in an undirected graph, an edge from a to b is the same as an edge from b to a. The edges may also have associated numeric values, typically called *weights*. (In a graph of cities and roads between them, for instance, edge weights might correspond to the speed limits on the roads.) The degree of a vertex in a network is defined as the total number of incoming and outgoing edges. Vertexes with high degree are often called "hubs." The degree distribution $P(k)$ gives the probability that a selected vertex has exactly k such edges. The statistics of $P(k)$ can be used to characterize a graph, as discussed below.

Graphs appear frequently in the analysis of complex systems. Elsewhere in biology, the genes in a regulatory network and the metabolic chemicals in a cell are often modeled as graphs. Maps of computers on the Internet, transistors on a silicon chip, and people in social groups (61) are also well suited to graph representations. The theory of complex networks (62), originating in the mathematics and physics community, has become increasingly popular as a means of analyzing protein interaction networks.

If we consider the proteins to be a vertex set V and their interactions to be an edge set E, we can model protein–protein interactions as a graph. Protein networks are often represented as undirected graphs where a connecting edge signifies a binding between two proteins.

Analysis of networks typically starts with classification of the network architecture; protein networks are no exception. Three common mathematical models for network architecture include random networks, scale-free networks, and hierarchical networks (63) (Fig. 7). In the random network model, it is assumed that a fixed number of nodes are connected to each other at random. The vertices in random networks have Poisson degree distributions. Most nodes have roughly the same number of edges. The average path length $\ell\sim\ln N$, where N is the total number of vertices.

Scale-free networks, by contrast, have a power-law degree distribution. In scale-free networks, the probability that a node has k links is $P(k)\sim k^{-\gamma}$, where γ is the degree exponent (64). Most nodes in scale-free networks have few incoming/outgoing edges, whereas a handful of hubs have many edges. Hubs can serve as gateways in terms of network flow because they are linked to many other nodes. The average path length in such networks $\ell\sim\ln(\ln N)$, so messages may propagate more quickly random networks.

A scale-free network can be parameterized via a model where the probability distribution of the number of edges k is described as

$$P(k)=ak^{-\gamma}. \tag{2}$$

Here, a is the proportionality constant and γ is the degree exponent (63). This construct results in a small number of network hubs (nodes that have many interactions) relative to the more common nodes that have few links.

Modularity, local clustering, and scale-free topology are jointly exhibited in many biological systems. Such systems can be observed as combinations of recursive clusters culminating in what is termed *a hierarchical network*. A hierarchical design is formed through the interconnection of sparsely connected nodes that are part of highly clustered areas. Communication between the different highly clustered neighborhoods is often mediated via hubs.

Studies have shown protein–protein interaction networks to be scale-free (65). Most proteins participate in only a few interactions, whereas a few hubs participate in many (61) (Fig. 8). Scale-free networks are vulnerable to a targeted attack on a hub. In protein interaction networks, it has been shown that knocking out high connected proteins can cause catastrophic (lethal to organism) system failure (66).

These models have been applied to protein–protein interaction maps. Two early works (67,68) on the use of high-throughput y2H approaches in identifying potential protein–protein interactions between yeast proteins resulted in the discovery of 183 and 692 protein–protein interactions, respectively. Another study on *C. elegans* (69) showed the utility of the y2H method in identifying 27 novel protein–protein interactions. In addition, it helped to provide functional annotation for approximately 100 uncharacterized gene products in the worm via mapping to orthologous clusters (70).

Another interesting area of current research is the probabilistic prediction of protein interaction networks

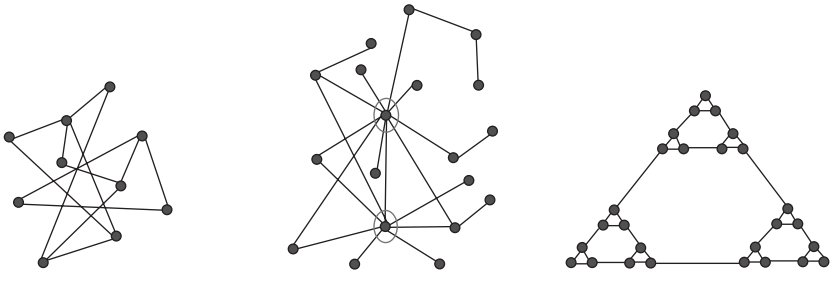

Figure 8. Subset of the yeast protein interaction network (YHR200W is a hub here).

(71). A study on yeast (72) correctly predicted the functional category for 72% of the 1393 characterized proteins with at least one partner of known function. This model has also been applied to predict functions of 364 previously uncharacterized proteins in yeast. Experimental data from various sources are used to construct a skeleton of known interactions, and statistical inference methods such as Bayesian networks can be used to predict or ascertain interactions between proteins.

Recent research efforts have mapped protein interactions into modules. Modules describe subgraphs (comprising interacting proteins) required for a specific cellular function. Understanding module function(s) requires direct knowledge of the involved proteins, where the proteins localized, and the module's regulation mechanisms. Integrating the information from different types of networks (metabolic, genomic, proteomic) can lead to a better understanding of functional modules (73). Researchers have also used protein–protein interaction maps to formulate new biological questions and hypothesis and for reducing problem complexity (74). The complete potential of protein interaction maps has yet to be exploited. Much promise resides in current interdisciplinary efforts aimed at mining the rich data contained within such networks.

5. CONCLUSION

The abundance, submicroscopic size, and dynamic nature of proteins have historically made them difficult to explore. On the other hand, these features also make proteins an ideal complex system for engineering-based analysis. Accurate sensors and signal processing methods are needed to rapidly assay protein abundance and interaction. High-throughput robotic systems are needed to 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 the scientist in analyzing experimental results and make whole-organism proteome-based experiments a reality.

Future research in proteomics will benefit from both new data-gathering technology and new data-processing methods. The field abounds with collaborative opportunities: Whereas the design of a protein chip or novel machine-learning algorithm may require skills in mechanical engineering or mathematics, the design of an experiment and the interpretation of its results will require the skills of a biologist or a biochemist. Innovative approaches, ranging from constructing accurate cellular models to building better detection instruments to formulating experimental hypotheses, will drive the future of proteomics. In this new era, proteomics is not merely validating hypotheses but generating new ones.

BIBLIOGRAPHY

1. V. K. Mootha, J. Bunkenborg, J. V. Olsen, et al., Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*. 2003; **115**(5):629–640.
2. M. R. Wilkins, From proteins to proteomes: large scale protein identification by two dimensional electrophoresis and amino acid analysis. *Biotechnology.* 1996; **14**:61–65.
3. D. D. Shoemaker and P. S. Linsley, Recent developments in DNA microarrays. *Curr. Opin. Microbiolo*. 2002; **5**:334–337.
4. M. F. Templin, D. Stoll, J. Bachmann, et al., Protein microarrays and multiplexed sandwich immunoassays: What beats the beads? *Comb. Chem. High Throughput Screen*. 2004; **7**(3):223–229.
5. P. Najmabadi, A. A. Goldenberg, and A. Emili, Conceptual design for an automated high-throughput magnetic protein complex purification workcell. *JALA*. 2003; **8**(6):101–106.
6. P. Bertone and M. Gerstein, Integrative data mining: The new direction in bioinformatics. *IEEE Eng. Med. Biol. Mag*. 2001; **20**(4):33–40.
7. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of The Cell*. New York: Garland Science, 2002.
8. C.-I. Branden and J. Tooze, *Introduction to Protein Structure*. 2nd ed. New York: Garland Publishing, 1999.
9. D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, Protein function in the post-genomic era. *Nature*. 2000; **405**:823–826.
10. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science*. 1998; **282**(5396):2012–2018.
11. M. M. Ruedi Aebersold, Mass spectrometry-based proteomics. *Nature*. 2003; **422**:198–207.
12. B. Bogdanov, R. D. Smith, Proteomics by FTICR mass spectrometry: Top down and bottom up. *Mass Spectrom. Rev.* 2005; **24**(2).
13. G. Alterovitz, A Bayesian Framework for Statistical Signal Processing and Knowledge Discovery in Proteomic Engineering. In: *Electrical and Biomedical Engineering*. Cambridge, MA: Massachusetts Institute of Technology, 2005, p. 89.
14. G. A. Michaud and M. Snyder, Proteomic approaches for the global analysis of proteins. *Biotechniques*. 2002; **33**:1308–1316.
15. D. C. Leibler, *Introduction to Proteomics: Tools for the New Biology.* Totowa, NJ: Humana Press, 2002.
16. J. Lill, Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom. Rev.* 2003; **22**:182–194.
17. C. Hoog, Proteomics. *Annual Rev. Genomics Human Genet*. 2004; **5**:267–293.
18. H. Steen and M. Mann, The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol*. 2004; **5**(9):699–711.
19. M. Yarmush and A. Jayaraman, Advances in proteomic technologies. *Annu. Rev. Biomed. Eng*. 2002; **4**:349–373.
20. S. Mitra and R. Brukh, *Sample Preparation Techniques in Analytical Chemistry.* New York: Wiley, 2003.
21. M. Yarmush and A. Jayaraman, Advances in proteomic technologies. *Annu. Rev. Biomed. Eng*. 2002; **4**:349–373.
22. J. R. Yates, Mass spectrometry and the age of the proteome. *J. Mass Spectrom*. 1998; **33**:1–19.
23. M. Merchant and S. R. Weinberger, Recent advancements in surface enhanced lased desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*. 2000; **21**:1164–1167.
24. E. F. Petricoin, A. M. Ardekani, B. A. Hitt, et al., Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002; **359**(9306):572–577.

25. E. F. Petricoin, K. C. Zoon, E. C. Kohn, et al., Clinical proteomics: Translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* 2002; **1**(9):683–695.
26. S. Hanash, Disease proteomics. *Nature* 2003; **422**:226–232.
27. G. R. E. A. Dalmasso, SELDI ProteinChip array technology: Protein-based predictive medicine and drug discovery applications. *J. Biomed. Biotechnol*. 2003; **4**:237–241.
28. L. Zhang, Contribution of human α-defensin 1, 2 and 3 to the anti-HIV activity of CD8 antiviral factor. *Science* 2002; **298**:995–1000.
29. G. Alterovitz, M. Aivado, D. Spentzos, et al., Analysis and robot pipelined automation for SELDI-TOF mass spectrometry. Proc. Int. Conf. of IEEE Engineering in Medicine and Biology, San Francisco, CA, 2004.
30. A. G. Marshall, Ion cyclotron resonance and nuclear magnetic resonance spectroscopies: Magnetic partners for elucidation of molecular structure and reactivity. *Acc. Chem. Res*. 1996; **29**(7):307–316
31. H.-W. Lahm and H. Langen Mass spectrometry: A tool for the identification of proteins separated by gels. *Electrophoresis* 2000; **21**:2105–2114.
32. M. Mann and H. R. Pandey, Analysis of proteins and proteomes by mass spectrometry. *Annual. Rev. Biochem*. 2001; **70**:437–473.
33. B. Kuster, J. S. Andersen, and M. Mann, Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 2001; **1**:641–650.
34. D. J. Pappin and A. J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol*. 1993; **3**:327–332.
35. W. H. Tang, B. R. Halpern, I. V. Shilov et al., Discovering known and unanticipated protein modifications using MS/MS database searching. *Anal. Chem*. 2005; **77**(13):3931–3946.
36. M. Quadroni, Proteomics and automation. *Electrophoresis* 1999; **20**:664–677.
37. J.R. Yates, Database searching using mass spectrometry data. *Electrophoresis*. 1998; **19**:893–900.
38. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; **18**:3551–3567.
39. W. Zhang and B. T. Chait, ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem*. 2000; **72**(11):2482–2489.
40. D. L. Tabb, A. Saraf, and J. R. Yates 3rd, GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem*. 2003; **75**(23):6415–6421.
41. J. R. Yates, J. K. Eng, A. L. McCormack, and D. Schieltz, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem*. 1995; **67**:1426–1436.
42. S. P. Gygi, B. R. Franza, and R. Aebersold, Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol*. 1999; **19**:1720–1730.
43. B. B. Haab, M. J. Dunham, and P. O. Brown, Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol*. 2001; **2**:RESEARCH0004.
44. N. Grebenchtchikov, S. P. Van Broekhoven, D. De Jong, A. Geurts-Moespot, P. N. Span, H. A. Peters, H. Portengen, J. A. Foekens, C. G. Sweep, and L. C. Dorssers, Development of an ELISA for measurement of BCAR1 protein in human breast cancer tissue. *Clin. Chem*. 2004; **50**:1356–1363.
45. P. Uetz and A. Grigorive, The yeast interactome. In: M. J. Dunn et al., eds., *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. New York: Wiley, 2005.
46. L. V. Zhang, S. L. Wong, O. D. King et al., Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinform*. 2004; **5**:38.
47. D. L. Wheeler, T. Barrett, D. A. Benson et al., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2005; **33**(Database issue):D39–D45.
48. C. Brooksbank, G. Cameron, and J. Thornton, The European Bioinformatics Institute's data resources: Towards systems biology. *Nucleic Acids Res*. 2005; **33**(Database issue):D46–D53.
49. P. E. Bourne, J. Westbrook, and H. M. Berman, The Protein Data Bank and lessons in data management. *Brief Bioinform*. 2004; **5**(1):23–30.
50. I. Xenarios, L. Salwinski, X. J. Duan et al., DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*. 2002; **30**(1):303–305.
51. C. Alfarano, C. E. Andrade, K. Anthony et al., The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*. 2005; **33**(Database issue):D418–D424.
52. M. Kanehisa and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000; **28**(1):27–30.
53. I. M. Keseler, J. Collado-Vides, S. Gama-Castro et al., EcoCyc: A comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*. 2005; **33**(Database issue):D334–D337.
54. H. Hermjakob, L. Montecchi-Palazz, G. Bader et al., The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol*. 2004; **22**:177–183.
55. A. K. Ramani, R. C. Bunescu, R. J. Mooney et al., Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*. 2005; **6**(5):R40.
56. J. T. Prince, M. W. Carlson, R. Wang et al., The need for a public proteomics repository. *Nat. Biotechnol*. 2004; **22**(4):471–472.
57. K. Michalickova, G. D. Bader, M. Dumontier et al., SeqHound: Biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinform*. 2002; **3**(1):32.
58. F. Iragne, A. Barre, N. Goffard et al., AliasServer: A web server to handle multiple aliases used to refer to proteins. *Bioinformatics* 2004; **20**(14):2331–2332.
59. O. Poetz, R. Ostendorp, B. Brocks et al., Protein microarrays for antibody profiling: Specificity and affinity determination on a chip. *Proteomics* 2005; **5**(9):2402–2411.
60. P. Legrain and L. Selig, Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett*. 2000; **480**(1):32–36.
61. A. Barabasi and Z. Oltvai, Network biology: Understanding the cell's functional organization. *Nature Rev. Genet*. 2004; **5**:101–113.
62. R. Albert and A. L. Barabasi, Statistical mechanics of complex networks. *Rev. Modern Phys*. 2002; **74**:47–97.
63. A. L. Barabasi and Z. N. Oltvai, Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet*. 2004; **5**(2):101–113.
64. A. Barabasi and R. Albert, Emergence of scaling in random networks. *Science* 1999; **286**:509–512.

65. H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, Lethality and centrality in protein networks. *Nature* 2001; **411**:41–42.

66. H. Jeong, S. P. Mason, A. L. Barabasi et al., Lethality and centrality in protein networks. *Nature* 2001; **411**(6833):41–42.

67. P. Uetz, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000; **403**:623–627.

68. T. Ito, K. Tashiro, S. Muta et al., Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA*. 2000; **97**(3):1143–1147.

69. A. J. M. Walhout et al., Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 2000; **287**:116–122.

70. R. L. Tatusov, E. V. Koonin, and D. J. Lipman, A genomic perspective on protein families. *Science* 1997; **278**(5338):631–637.

71. G. A. Petsko, Analyzing Molecular Interactions. In: *Current Protocols in Bioinformatics*. New York: Wiley, 2003.

72. B. Schwikowski, P. Uetz, and S. Fields, A network of protein–protein interactions in yeast. *Nature* 2000; **18**:1257–1261.

73. S. Tornow and H. W. Mewes, Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*. 2003; **31**:6283–6289.

74. S. J. Boulton, S. Vincent, and M. Vidal, Use of protein interaction maps to formulate biological questions. *Curr. Opin. Chem. Biol*. 2001; **5**:57–62.