Mass spectrometry-based proteomics and analyses of serum: A primer for the clinical investigator

V.A. Fusaro[1], J.H. Stone[2]

[1]National Cancer Institute/Food & Drug Administration, Clinical Proteomics Program; [2]Division of Rheumatology, Johns Hopkins University School of Medicine

Vincent A. Fusaro, BS, Computer Scientist, [1]National Cancer Institute/Food & Drug Administration, Clinical Proteomics Program; John H. Stone, MD, MPH, Associate Professor of Medicine, Director, The Johns Hopkins Vasculitis Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

Please direct correspondence and reprint requests to: John H. Stone, MD, MPH,
The Johns Hopkins Vasculitis Center,
5501 Hopkins Bayview Circle, Baltimore,
Maryland 21224, USA.
E-mail: jstone@jhmi.edu

Received on August 12, 2003; accepted in revised form on September 27, 2003.

Clin Exp Rheumatol 2003; 21 (Suppl. 32): S3-S14.

Key words: Proteomics, vasculitis, mass spectrometry.

ABSTRACT

The vocabulary of proteomics and the swiftly-developing, technological nature of the field constitute substantial barriers to clinical investigators. In recent years, mass spectrometry has emerged as the most promising technique in this field. The purpose of this review is to introduce the field of mass spectrometry-based proteomics to clinical investigators, to explain many of the relevant terms, to introduce the equipment employed in this field, and to outline approaches to asking clinical questions using a proteomic approach. Examples of clinical applications of proteomic techniques are provided from the fields of cancer and vasculitis research, with an emphasis on a pattern recognition approach.

Sir William Osler (1912) (1)

"In the capillary lake into which the arterial stream widens, the current slows and the pressure lessens … In the brief fraction of a second … the business of life is transacted, for here is the mart or exchange in which the raw and the manufactured articles from the intestinal and hepatic shops are spread out for sale".

Introduction

We live in both a remarkable period in the history of science and a time of unprecedented opportunity in clinical investigation. As the quote from Osler indicates, clinicians have long recognized that the critical mechanisms of both health and disease play themselves out at the level of the microvasculature. The essential "manufactured articles" to which Osler refers were recognized – even in his day – as proteins. As the genetic code's effectors, proteins determine the phenotype not only of each cell, but also of every tissue and organ (and ultimately the entire organism). Although the concept that "Genes are destiny" is true with regard to some disorders, in many others the identification of candidate genes has revealed disappointingly little about disease biology, patients' response to therapy, the impact of lifestyle changes on health, and other important issues. Such types of information are reflected more reliably in the levels of mRNA and even more so in the specific types and quantities of the proteins themselves that are expressed (Fig. 1).

In theory all disease processes, even those based in single organs, lead to

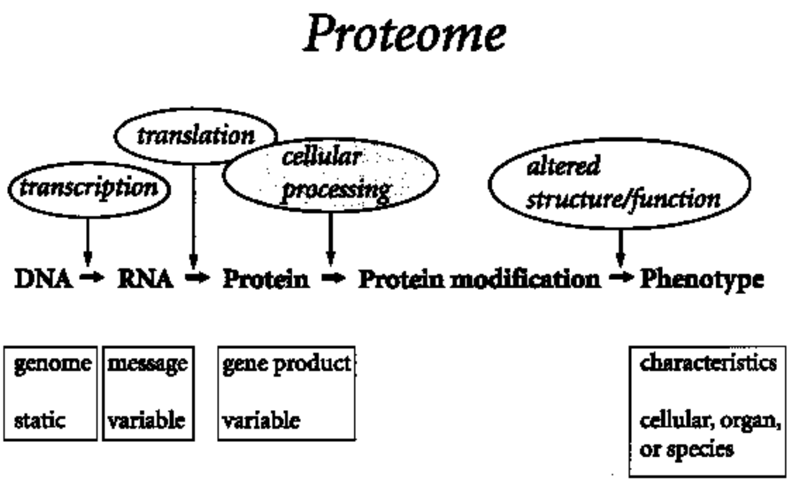

Fig. 1. Diagram of assorted cellular processes leading to phenotype and the relationship of these processes to proteomics.

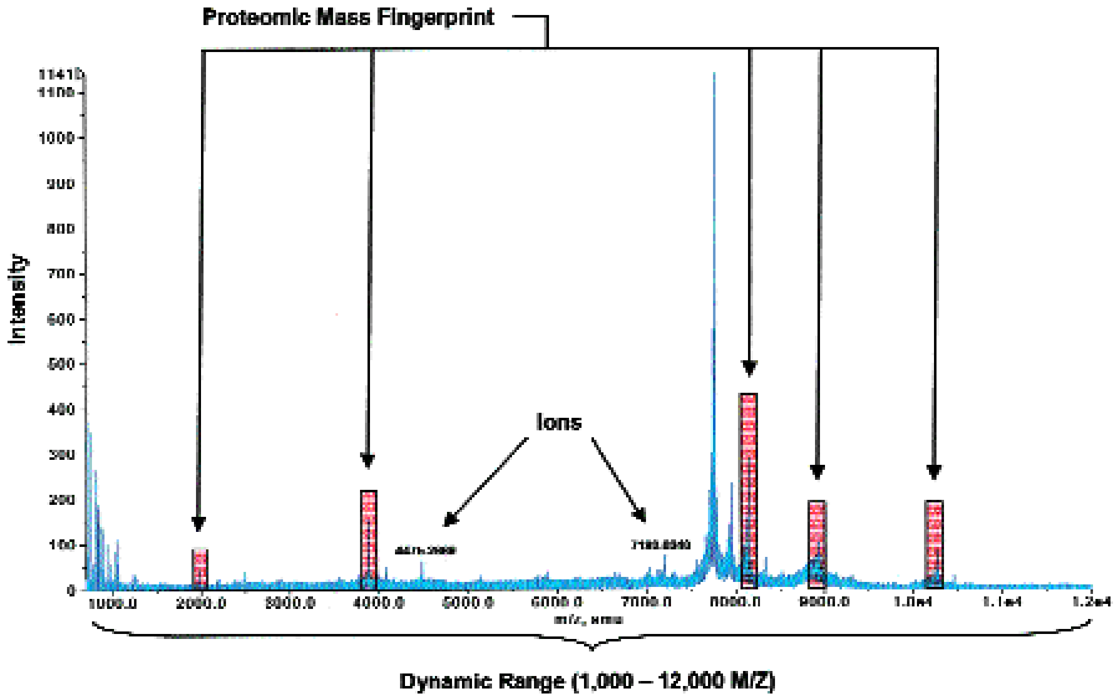

Fig. 2. Prototypical proteomic profile illustrating several glossary terms. The profile is the readout from a single patient sample analyzed by tandem mass spectrometry.

perturbations within the serum. As discussed in this article, proteomic studies in ovarian and prostate cancer support this concept (2,3). The application of proteomic techniques to human serum may also have particular relevance to inflammatory vascular conditions such as systemic vasculitis, a group of disorders in which the site of pathology – the blood vessel wall – is in direct contact with the serum. The ability to make accurate inferences about the state of pathology (or health) within organ systems by examining the fluid that perfuses them has several major potential advantages. First, because traces of the molecular footprints of disease are expected to equilibrate (even at subminute quantities) in the serum, the strong possibility of sampling error that often accompanies tissue biopsy is reduced substantially. Second, findings in the serum represent the sum of disease processes in organs, even those in which clinical involvement is unrecognized. This may be particularly relevant to multi-organ system diseases such as vasculitis. Finally, because phlebotomy can be repeated essentially as often as needed, serum investigations provide relative ease of sampling compared to the biopsy of solid organs.

The “Proteome” and “Proteomics”: Working definitions

The words “proteome” and “proteomics” did not even exist ten years ago. One may guess, from the burgeoning of terms that end in “-omics”, that proteomics is the study of the “proteome”. But what does this term mean? A proteome may be considered to be all of the **prote**ins linked to a given set of genes (gen**ome**). These proteins include not only those translated directly from genes but also those modified after translation. All proteins present in a cell or organism at a given time comprise its proteome. Moreover, investigators also refer to “subproteomes”, which may be restricted to specific biological compartments, e.g., the inner mitochondrial membrane. *Proteomics* is the application of tools from fields as diverse as clinical medicine, molecular biology, mass spectrometry, and bioinformatics to explore the separation, identification, and characterization of proteins, and to shape this wealth of information into new knowledge. Thus, proteomics is not a single discipline but rather a collection of highly-specialized forms of expertise, all of which may be brought to bear on many types of clinical problems.

Glossary of major terms

We present below a glossary of terms for which the meanings are not intuitively clear, despite their frequent use in the proteomics literature. Beginning with this glossary will orient the reader, even if the context of all the terms is not apparent initially. The definitions of the terms included often contain other terms that are defined elsewhere in the glossary. These other terms are highlighted in **bold**. Figure 2 illustrates several of the glossary terms and other concepts in this review.

Abundance: The proteomics literature refers to “low abundance” proteins (e.g., cytokines) and “high abundance” proteins (e.g., albumin or immunoglobulins). The term abundance simply means *concentration*. The development of methods by which low abundance proteins may be studied in the setting of high abundance proteins that dwarf them by many orders of magnitude is one of the greatest conundrums confronting proteomics today.

Analytes: Proteins and peptides contained within the clinical sample to be analyzed.

Dynamic range: Refers to the proteins and peptides within the proteome that are defined by a set of certain charac-

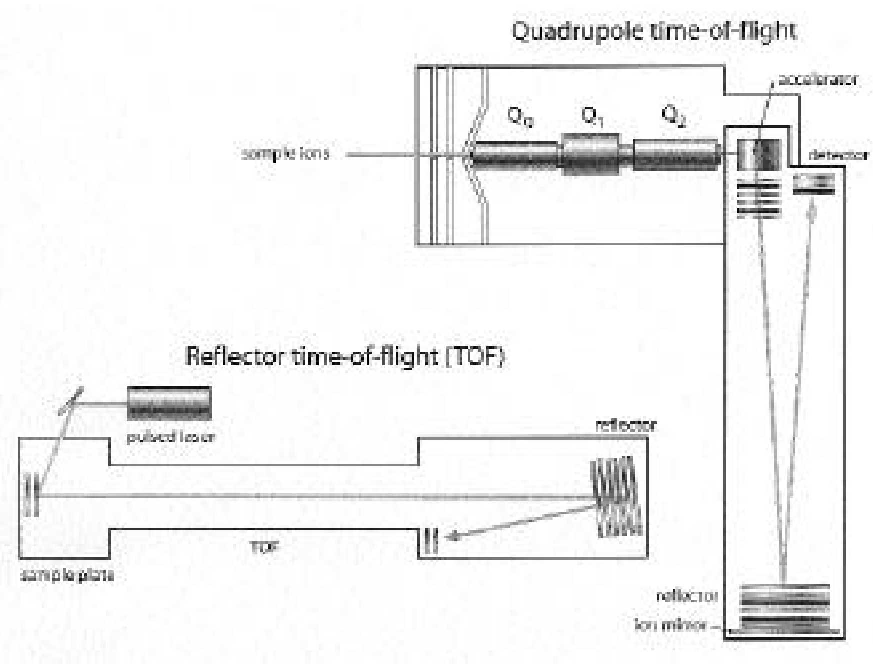

Fig. 3. Schematic diagrams of mass spectrometers.
A. Standard low-resolution mass spectrometer, typical of those used in MALDI and SELDI analyses.
B. Tandem mass spectrometer. The principal distinguishing feature from low-resolution instruments is the presence of a collision cell in which peptides are broken apart by collision with an inert gas, permitting in many cases sequence analysis and parent protein identification.

teristics, e.g., molecular weight (MW), charge, abundance, or other features. Approaches to proteomic analyses may be evaluated in part by the dynamic range of the proteome to which they provide access. With regard to MW, for example, some proteomic techniques (see **SELDI**) that are highly effective in analyses of ions and peptides in the range of 700-12,000 Daltons (Da) may have little utility at MW beyond this range.

Electrospray ionization (ESI): A technique commonly used to volatilize and ionize proteins or peptides for mass spectrometric analysis. ESI ionizes **analytes** out of solution. It is readily coupled, therefore, to liquid-based protein separation tools such as liquid chromatography (LC). Integrated systems of LC and mass spectrometry (LC-MS), now based on ESI, are the preferred technique for the analyzing complex samples.

Fractionation: A step in the preparation of samples for some types of proteomic analysis. Fractionation involves the use of a variety of techniques to remove certain proteins from a sample (e.g., high **abundance** proteins). In some cases, the removal of “interference” by such proteins facilitates the analysis of other proteins of interest (e.g., those of lower abundance) that are theoretically more pertinent to the disease of interest. *Fractionation* must be distinguished from *separation*, which is the differentiation of proteins and peptides from each other that usually occurs (by mass spectrometry or another technique) after fractionation has been performed.

Intensity: Refers to the height of a peak at a given mass-to-charge ratio in a proteomic profile (Fig. 2). The intensity is the number of times an **ion** of a particular **mass:charge ratio** strikes the analyte detector during a data acquisition period. An important (but counterintuitive) point is that the intensity of a given peak correlates poorly with the quantity of ion in a specimen.

Ions: Strictly speaking, in mass spectrometry the **analytes** are typically ions (charged particles) rather than full proteins or peptide fragments. In their ionized state, analytes may be separated by the mass spectrometer on the basis of their **mass:charge ratios**.

MALDI: An abbreviation for **m**atrix-**a**ssisted **l**aser **d**esorption/**i**onization, a traditional platform for mass spectrometry. MALDI consists of a stainless steel plate onto which the sample is spotted directly.With the MALDI technique, **analytes** are sublimated (i.e., taken directly from the solid to the gaseous phase) and ionized out of a dry, crystalline matrix by laser pulses. Protein separation using affinity columns or other **fractionation** techniques is usually performed before the application of mass spectrometry by MALDI.

Mass-to-charge ratio: (Abbreviated *m/z*). The ratio of the mass of an ionized peptide or protein to its overall charge. The *m/z* ratio comprises the X-axis (Fig. 2) on the output of proteomic spectra from mass spectrometers, and corresponds loosely to MW. Mass spectrometrists often speak of an ion’s "mass", when technically they are referring to its *m/z* ratio. This convenient way of speaking is really only accurate in the case of singly-charged ions.

Mass spectrometry (MS): (Fig. 3A and B) An instrument that measures the masses of individual molecules that have been converted to ions, i.e. that are electrically charged. In general, a mass spectrometer has three components: 1) a chamber that holds the ion source (the clinical sample from which **analytes** are ionized via laser); 2) a detector that registers the number of ions at each *m/z* value; and, 3) a mass analyzer that measures the *m/z* ratio of the ionized analytes. A mass spectrometer has the ability to analyze samples processed on a variety of platforms, including **MALDI**, **SELDI**, and **ESI**.

Peptide mass mapping: One method of identifying proteins whose masses have been determined by MS. The identity of proteins is established by matching the analytes’calculated masses with the lists of all peptide masses at entries in publicly-accessible databases (e.g., SWISS-PROT or TrEMBL). A more “cutting-edge” method of protein identification, which exploits the capabilities of **tandem MS** (Fig. 3B), is the analysis of collision-induced spectra, described below. Because neither peptide mass mapping nor tandem MS is capable of identifying all peptides or proteins, the two approaches are com-

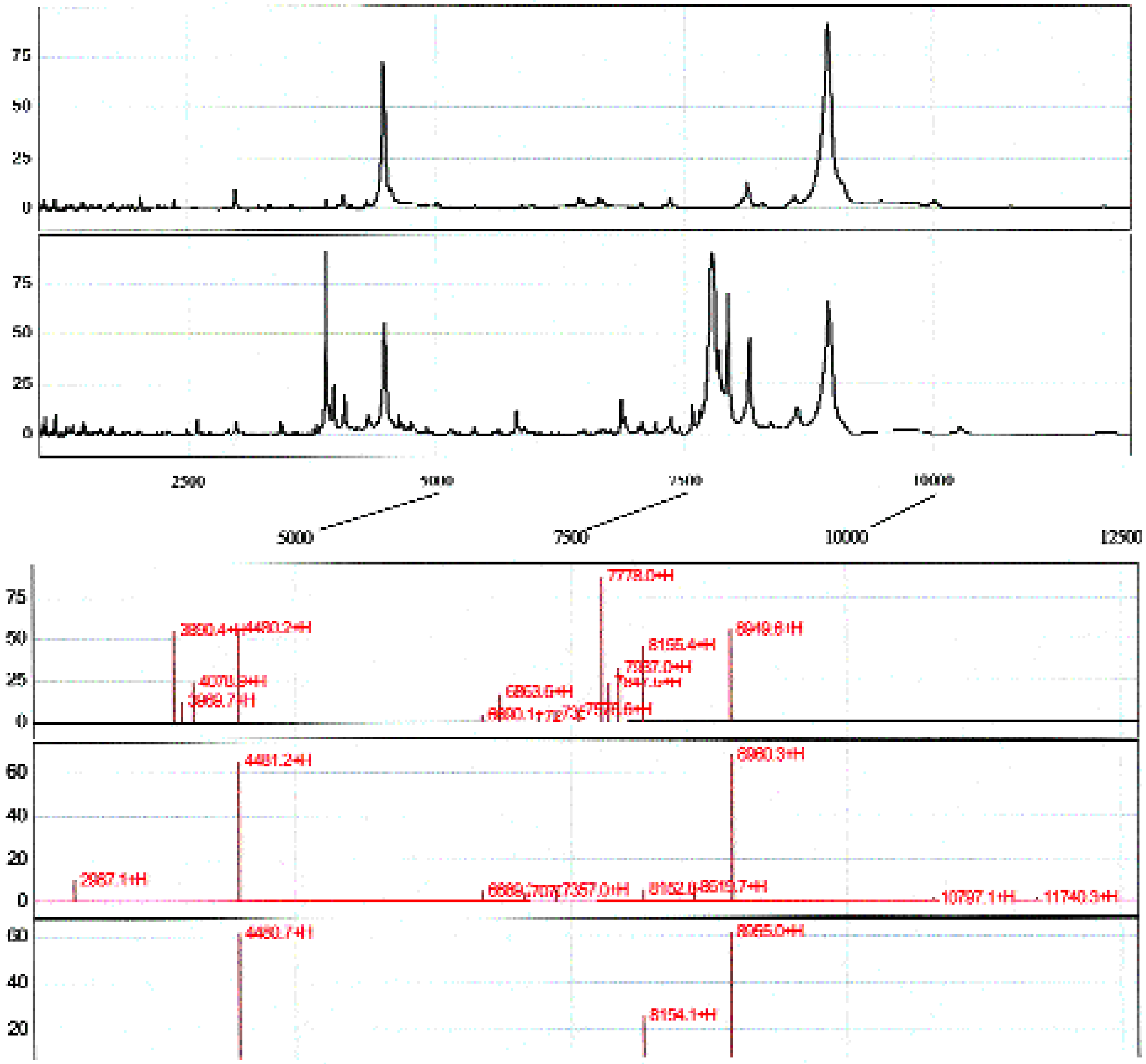

Fig. 4. Peak map (lower part of the figure, in red) showing how numerous analytes may be found around individual *m/z* values.

plementary.

Proteomic mass fingerprint (PMF): Refers to a unique combination of ions whose overall intensity differences can segregate different states (e.g., samples from patients with cancer from those of patients who do not have cancer). As discussed below, a PMF consisting of 5 ions has been shown to discriminate patients who have ovarian cancer from those who are at high risk but who are cancer-free (2).

Resolution: Refers to the ability of a mass spectrometer (or, more specifically, of its mass analyzer) to distinguish between discrete analytes with similar characteristics (see, for example, the peak map in Fig. 4). In general, tandem MS techniques have greater resolution than their predecessors, albeit their **dynamic ranges** may be considerably narrower.

SELDI: Abbreviation for surface-enhanced laser desorption/ionization, another type of platform for MS studies. The SELDI technique performs protein separation based on the analytes' surface charge. First applied to clinical medicine in the late 1990s, SELDI represents a breakthrough in protein separation techniques because of its superiority (compared to two-dimensional gel electropheresis, 2-DE) in the detection of low MW ions and ions of basic charge.

Tandem mass spectrometry: Also referred to as MS/MS (and, when coupled to liquid chromatography, as LC-MS/MS). Ions of a particular *m/z* value are selected by a first mass analyzer, and then fragmented in a collision cell (Fig. 3B). The masses of the ion fragments are subsequently "read out" by a second **time-of-flight** mass analyzer. A sequence as short as 5 amino acid residues may be sufficient to identify an entire protein provided that the sequence is not derived from a highly-conserved motif.

Time-of-flight: Refers to the length of time required for proteins and peptides ionized from the surface of a protein chip to travel through the MS chamber to the detector plate. *Time of flight* is

Table I. Functional groups of blood proteins*.

Proteins secreted by solid tissues that act in serum
* Largely produced in the intestines and liver
* Include the classic serum proteins (e.g., albumin)

Immunoglobulin

“Long-distance” receptor ligands
* Classic peptide and protein hormones (e.g., erythropoietin and insulin)

“Local” receptor ligands
* Cytokines
* Mediate local interactions and are subsequently diluted into serum at ineffective levels
* Native MWusually < kidney filtration cut-off

Temporary passengers
* Non-hormone proteins that traverse the serum transiently en route to the site of their primary function (e.g., proteins secreted elsewhere but sequestered in lysosomes)

Tissue leakage products
* Released into serum as a result of cell death/damage (e.g., troponin, creatine kinase)

Aberrant secretions
* Tumor-associated proteins or secretions from other abnormal tissues

Foreign proteins
* Proteins related to infectious agents

*Adapted from (21).

abbreviated “TOF”, as in “SELDI-TOF” or “MALDI-TOF” or “quadrupole-TOF (Q-TOF)”. The fundamental principle that permits MS to separate analytes is the fact that small ions fly faster than large ones. The ions’ *m/z* ratios may be calculated from the time that each requires to reach the detector plate. Differences in TOF permit the distinction and, in many cases, the identification (by tandem MS) of different peptides.

The proteome’s inherent challenges

Examinations of a blood substance called “albumin” began as early as the 1830s (4). Thus, in some ways only the name “proteomics” is new. Our appreciation of the depth and complexity of the proteome continues to evolve with the development of new techniques for studying it. Studying the proteome has numerous challenges, some inherent to the nature of protein mixtures themselves and others more specific to the potential application (i.e., the disease or clinical question of interest).

Complexity

Until very recently, the concept of “one gene, one protein” was regarded as fundamental to biology.We now recognize that this concept is a remarkable underestimation of the proteome’s complexity. Because of splicing, processing, post-translational modifications, and other events that occur once proteins have been made, the proteome is considerably more complex than the genome. In contrast to the 30,000 – 50,000 genes that comprise the human genome, serum probably contains millions of polypeptide species, spanning a staggering concentration range of 10 orders of magnitude. A sobering fact today is that even with the most robust MS techniques, only about 500 proteins have been identified to date (5). (These include many of the proteins used today in clinical evaluations, e.g., creatine kinase, troponin, and aspartate and alanine aminotransferase). A list of the broad functional groups of blood proteins known currently is shown in Table I. Contrary to the status of the human genome, a full description of the human proteome is a task for which completion is not even nearly in sight. Moreover, in contrast to the “shotgun” sequencing approach that permitted the rapid completion of the Human Genome Project, there is not yet a clear road map or set of techniques for mapping the entire proteome.

Protein binding

More than half of the known proteins are smaller than the presumed size cut-off of the glomerular filtration apparatus (approximately 45 kDa). In theory, proteins below this MW should be lost in the urine on the first pass through the circulation. In order to remain in the serum, these proteins must either be part of larger protein complexes or possess other retention mechanisms. One likely explanation is that many low MW proteins are bound to albumin and/or other high abundance proteins. Simple stoichiometry dictates that most small, low abundance peptides within the serum will be bound to larger, charged species that are far more numerous. This point has profound implications for any attempts to fractionate serum specimens, simply because the removal of high abundance proteins almost certainly means that lower abundance proteins (peptides) are removed as well.

Dynamic range

Among the high abundance proteins, serum albumin has a concentration of 35-50 pg/ml. This single protein accounts for fully 55-60% of all proteins within the serum. In contrast, at the low abundance end the concentration of interleukin-6 – between 1 and 5 pg/ml – is 10^{-10} smaller. Comparing the concentration of IL-6 in the circulation to that of albumin is analogous to comparing the mass of a single human being to the combined mass of the entire human population (now nearly 7 billion people). The task of measuring masses across such an enormous concentration range constitutes one of the major challenges to complete description of the proteome. Even LC/MS/MS, the most versatile method for unbiased protein discovery, has a maximal dynamic range of only 10^4. Independent fractionation methods expand the possible dynamic range by only an additional 10^2 or so.

Timing

The proteome is vibrant, changing continually in response to its environment even after removal of a clinical specimen from a patient. In order to study the proteome accurately therefore, samples must be processed in a swift and uniform fashion. Failure to process samples quickly permits ongoing post-translational modifications to alter their detected phenotypes. MS can detect differences between samples processed immediately and those stored overnight in a refrigerator before processing, and also between samples that have been thawed only once before analysis and those that have been thawed several times.

Volume of data

Some approaches to the challenge of large quantities of data produced by proteomic analyses are outlined in the discussion of **Bioinformatics**, below.

Disease-specific challenges

Each individual disease poses its own set of challenges in the design of proteomic studies. Some of these challenges may include: 1) disease heterogeneity; 2) recognition of different stages of disease; 3) the collection of sufficient numbers of patients to provide adequate statistical power; 4) the timing of sampling with regard to disease activity; and 5) the impact of treatment on proteomic profiles. Just as technological limits create challenges to studying the proteome, the frequent lack of well-characterized clinical populations is another major hurdle that must be overcome before the promise proteomics can be realized. In the rush to embrace technology, there is a risk of overlooking the requirement for clean clinical phenotyping.

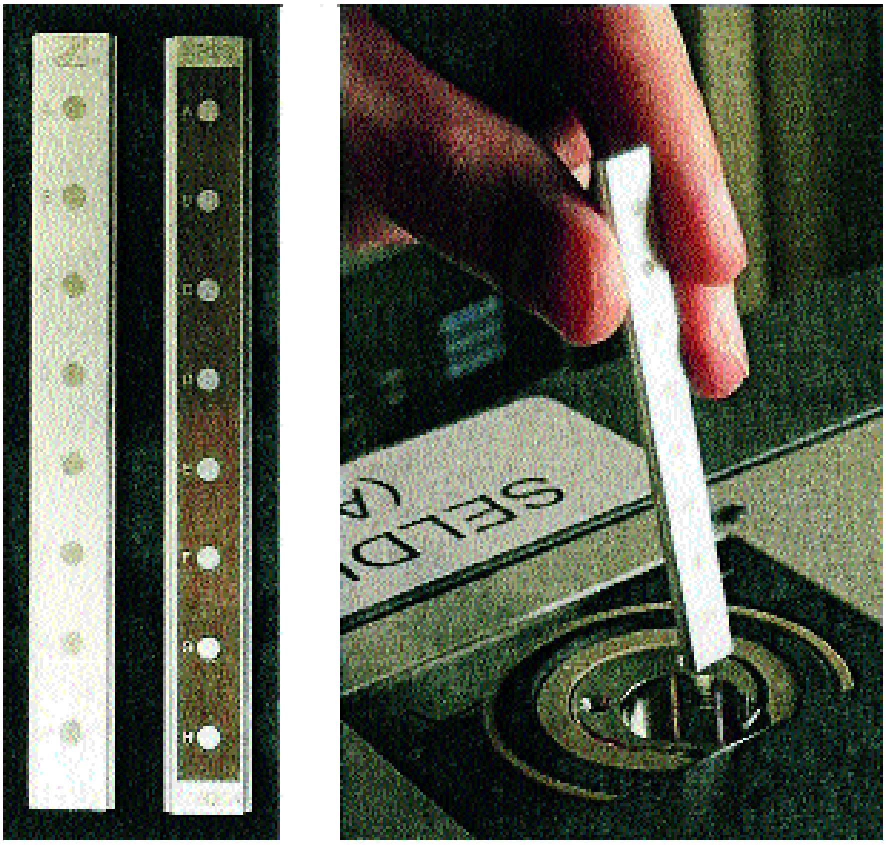
(A) (B)

Fig. 5. Protein chips. **A.** Two examples of protein chips. Each contains 8 spots (one spot for each clinical sample). The surfaces of the spots on the two types of chips shown contain different types of chip chemistries. **B.** Protein chip being inserted into a mass spectrometer for its encounter with the laser.

Limitations of older approaches to protein separation

For three decades, the mainstay of proteimic analysis has been 2-DE (6-8). In 2-DE, separation in the first dimension is achieved by isoelectric focusing according to the proteins' isoelectric point (pI). Proteins are then resolved orthogonally in the second dimension by their relative molecular mass, typically by SDS-PAGE. This approach has two major limitations as a tool for proteomics: firstly, 2-DE is ineffective at distinguishing low-abundance proteins; and secondly, 2-DE analyses underrepresent basic and membrane proteins. In

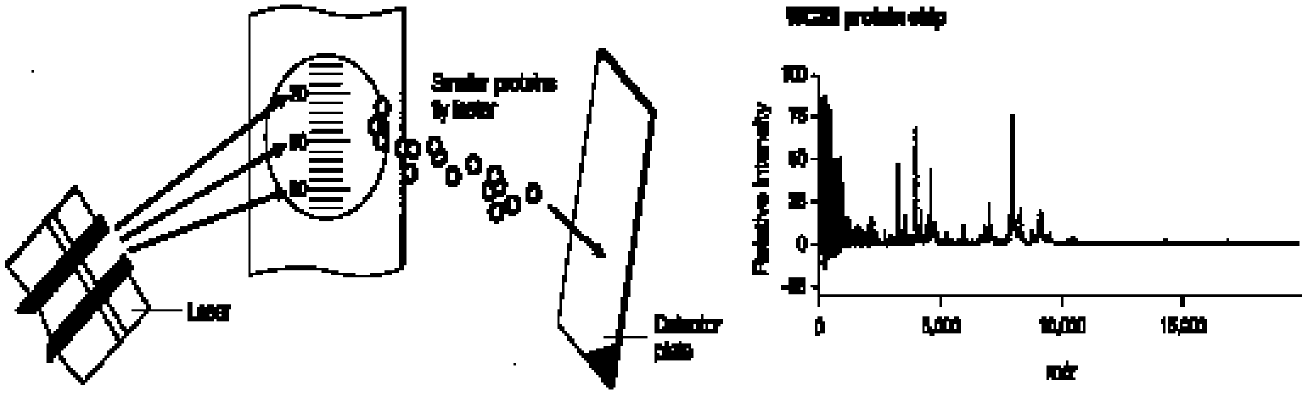

Fig. 6. Positions on protein chip spots. Each spot is divided into tiny coordinates (the 20, 50, and 80 marks shown in the figure) known as positions that are used to direct the laser to strike at precisely the same point on each spot. The right side of the figure shows the proteomic profile generated by the sample on the spot.

attempts to overcome these shortcomings, 2-DE analyses are now often coupled with MS technology; that is, spots of interest are selected, digested, and then analyzed by MS. Even so, this application has a limited dynamic range and is generally effective at the identification of only high abundance proteins. Because of its shortcomings, replacement of 2-DE has become in large measure the “Holy Grail” of proteomics. Although the disappearance of 2-DE is not likely to happen soon, developments of the past few years have made MS unrivaled for its accuracy in mass detection, its ability to address complex protein mixtures, its amenity to automation, and its high throughput capabilities.

Mass spectrometry: The basic components

The machines

With all MS instruments, peptides are ionized from samples using either a MALDI technique (from a solid state sample) or ESI (directly from the liquid phase). Generic MS instruments are depicted in Figures 3A and B. Most of the following discussion focuses on chip-based techniques of SELDI and tandem MS. The specific instrument to which we refer in the studies described below is the ABI Q-STAR Hybrid LC/MS/MS (Applied Biosystems; Foster City, CA) for SELDI processing. The upper limit of detection for this instrument is an *m/z* ratio of 12,000.

The chips

The SELDI protein chips are rectangular aluminum plates (Fig. 5) with approximate dimensions of 3" x 1/2" x 1/4". Each chip has 8 *spots* – one spot for each clinical sample. Furthermore, each spot has many positions (Fig. 6) that are not visible to the naked eye but which can be used to program the laser to strike precisely the same coordinates on each spot. There are many varieties of protein chips, each containing on their spot surfaces different substrates that are designed to target different dynamic ranges of peptides. Some substrates capture proteins with weakly positive charges, whereas others have affinities for metal ions such as nickel or copper. Because of the overlapping dynamic ranges that they target, different chip surfaces may be complementary. The chip essentially performs a protein separation on its surface, and samples can be pre-processed on the basis of size exclusion, pH, pI, and other features to further isolate proteins of interest. In general, the same protein chips used for SELDI analyses may also be used with tandem MS platforms.

Table II. Data from a hypothetical data mass spectrometry collection.

Acquisition	Ionic Species Detected		
	3000 *m/z*	5500 *m/z*	9800 *m/z*
Laser Fire #1	1	1	1
Laser Fire #2	1		
Laser Fire #3	1		1
Laser Fire #4	1	1	
Total Intensity	4	2	2

The matrix

In the processing of samples, a matrix is added to the chip surface after the application of the sample. The matrix forms a crystalline layer on top of the sample. The matrix crystals help transfer the laser energy to the sample, thereby aiding the ionization process and ultimately inducing the analytes to “fly” down the TOF tube. Without the addition of matrix, virtually no analytes become ionized.

The laser

Ionization occurs when energy is transferred from a laser beam to the sample. As noted, in the interests of sample-to-sample consistency, the laser pulses may be directed to precisely the same position on each spot. In the analysis of a clinical sample by MS, the laser may be fired at the sample thousands of times a second. Each firing of the laser at the sample and the resultant data on the *m/z* ratios of analytes are termed an “acquisition”.

The mass detector

Airborne ions strike a detector that records the presence of a “hit”. As shown in Table II, which contains hypothetical data, during the first laser pulse (acquisition #1) ionic species at *m/z* values of 3000, 5500 and 9800 hit the detector. During the second pulse, only a species with an *m/z* value of 3000 hit the detector; and so on. This type of data collection is multiplied and averaged for all of the ionized analytes from a given sample. The intensities of these ion species, ultimately reflected to some degree in the height of peaks on a proteomic profile (Fig. 2), are the sums of all the hits during the total acquisition time.

An important but counterintuitive point is that even tandem mass spectrometers are, at best, only semi-quantitative instruments. For both MALDI and ESI platforms, the relationship between the amount of a given analyte present and the measured signal intensity is complex and non-linear. The reasons for this phenomenon remain poorly understood.

Robots for sample processing

Swift advances in robotic instrumentation have led to tremendous increases in both the throughput capabilities and reproducibility of MS. Robots can be programmed to perform the entire chip preparation, including pre-treatment of the chip, sample application, and application of matrix. The advantages of robots include not only speed but also consistency in sample processing. Until recently, the typical time required to prepare 96 SELDI samples in the laboratory of the NCI-FDA Clinical Proteomics Program was approximately 3.5 to 4 hours. Even this representated a dramatic improvement over the time that would be required to process a comparable number of samples by hand. Recent breakthroughs in instrumentation have now decreased this time to 1.5 hours, and even greater throughput should be possible in the future.

Mass spectrometry platforms

MALDI is used most often for the analysis of comparatively simple peptide mixtures. Pharmaceutical companies have used the MALDI platform for years in the development of new drugs, specifically in the area of protein iden-

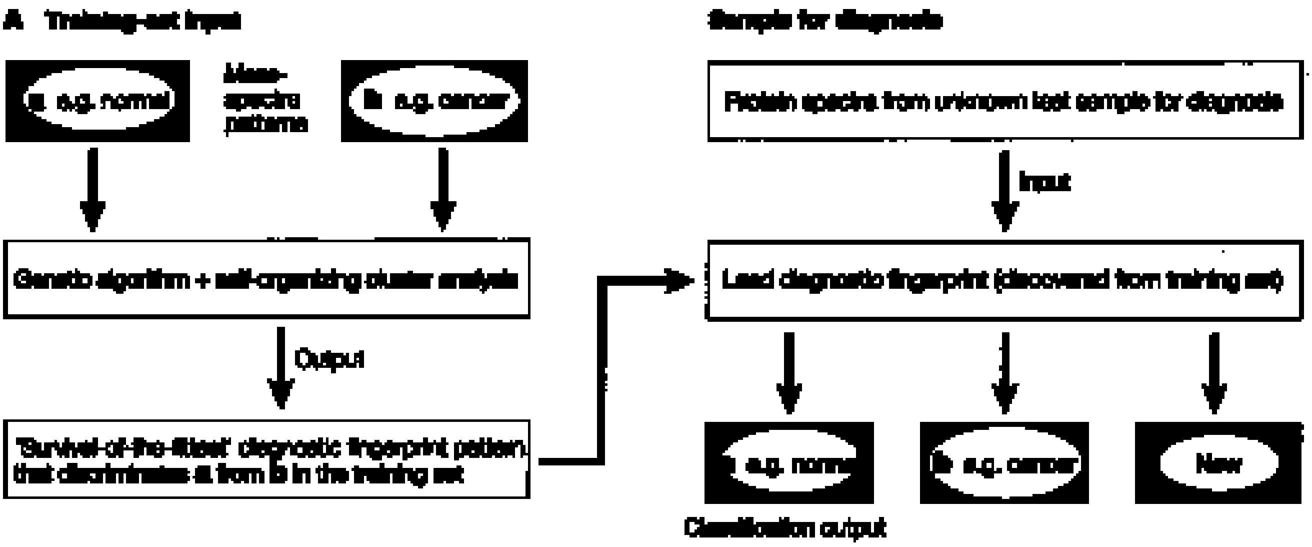
A Training-set input
a e.g. normal
Mass-spectra patterns
b e.g. cancer
Genetic algorithm + self-organizing cluster analysis
Output
'Survival-of-the-fittest' diagnostic fingerprint pattern that discriminates a from b in the training set
Sample for diagnosis
Protein spectra from unknown test sample for diagnosis
Input
Lead diagnostic fingerprint (discovered from training set)
a e.g. norma
b e.g. cancer
New
Classification output

Fig. 7. Schematic diagram of the approach to pattern recognition in proteomic studies.

tification. The standard procedure has been to query ion fragments identified against protein libraries to determine (when possible) the identity of their parent proteins.

The development of the SELDI platform offers several substantial advantages over two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (9-11). First, SELDI-TOF can describe entire populations of ions within serum simultaneously. Second, SELDI-TOF analysis is capable of detecting proteins that are smaller than 10,000 Daltons (Da), as well as proteins that are basically charged. The group of proteins in this lower MW range are of tremendous biologic potential because they contain cleaved or aberrantly shed proteins or peptides that may reflect essential features of a disease. Until recently, these molecules were below the level of detection (12).

A significant disadvantage of SELDI is that the technique does not provide a sequence-based identification, because there are many proteins close to a given *m/z* ratio (Fig.4). The protein peaks representing potential markers cannot be identified without significant additional effort. Tandem MS measurements now provide the means to characterize specific post-translational modifications and to identify structural differences between related proteins, differentially modified proteins, and protein isoforms (13-15). Individual proteins can be identified through the analysis of collision-induced spectra, which provide information about peptide sequences. Collision-induced spectra are scanned against comprehensive protein sequence databases (using a variety of possible algorithms). A peptide sequence tag approach identifies a short amino acid sequence from the peak pattern that, coupled with information about mass, permits determination of the peptide's origin (16). A technique known as "stable isotope labeling" now permits quantification of peptide levels by MS/MS (17, 18).

The profiling of protein "signatures"

The notion of a peptide mass fingerprint (PMF) has existed for several decades. In concept, the PMF is very simple: every disease will create characteristic changes within the proteome that permit the identification, staging, and other profiling of that specific disease. A disease's PMF may be used to differentiate that disorder from other diseases and from states of health. The combination of mass spectrometry and proteomics has become the method of choice for analyzing these differences. The technique described below possesses the advantage of not requiring fractionation and the consequent risk of removing low MW peptides of interest. All analytes within the dynamic range of the SELDI platform are potentially analyzable, provided that they become bound to the chip surface and ionized by the laser.

Disease-specific examples of functional proteomics

Ovarian and prostate cancer

Using SELDI-TOF analyses of sera, investigators have developed a method to distinguish the presence or absence of neoplasia within the ovary and prostate (2,3). These studies indicate that low-MW proteomic patterns exist in serum that reflect the pathologic state of the ovary and the prostate. Moreover, these patterns can predict the presence of ovarian cancer (including Stage I disease) and early prostate cancer with a high degree of reliability. Within the sera of these cancer patients, the use of novel bioinformatics techniques has identified optimal proteomic patterns that distinguish patients with these types of malignancies from relevant control groups. The flow diagram in Figure 7 provides an overview of this approach to proteomics. This approach is based on the simultaneous analysis of a pattern of proteins or peptide fragments, rather than reliance upon a pre-defined set of biomarkers.

The optimal discriminatory pattern identified for ovarian cancer consisted of relative abundances of proteins at 5 different MWs (534, 989, 2111, 2251, and 2465 Da) (2). In contrast, the optimal discriminatory pattern identified for prostate cancer consisted of relative

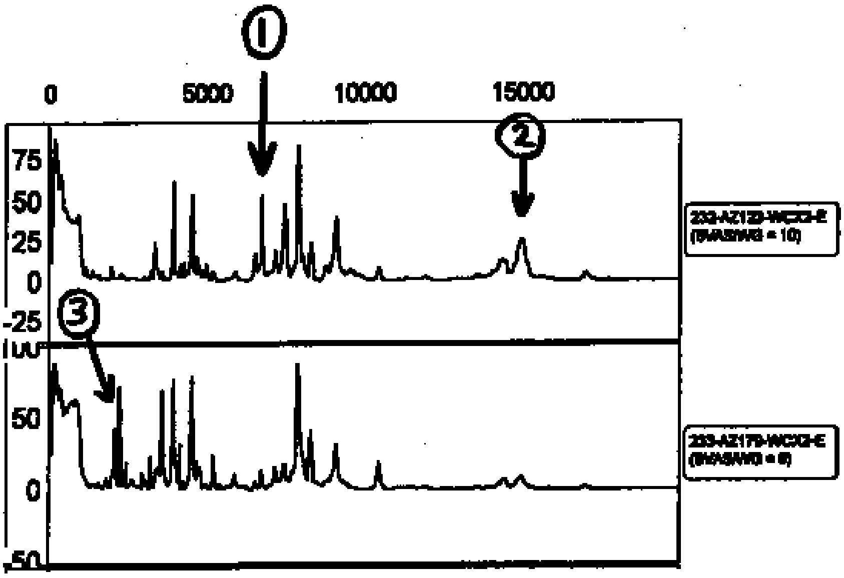

Fig. 8. Serum proteomic profiles from baseline and remission in a patient with Wegener's granulomatosis. Even on visual inspection, major differences between these two profiles are evident. Note the narrow band of intensity around the *m/z* value 7,000 in the baseline sample that is absent at remission. There are also two peaks just below 15,000 which are substantially more intense at baseline than in the remission sample. Finally, in the remission sample there is a broad range of intensity in the 2,500 – 3,000 range that is absent in the baseline sample.

abundances of proteins at 7 different MWs (2092, 2367, 2582, 3080, 4819, 5439, and 18220 Da), all distinct from those that segregated ovarian cancer. Most strikingly, the serum proteomic analyses were able to make the critical distinction between two different types of pathology in the prostate: frank prostate cancer and benign prostatic hypertrophy (3). Confirmation of this approach using a tandem MS platform is now being performed in the context of a multi-center clinical study.

Systemic vasculitis: Wegener's granulomatosis

Preliminary work indicates that these techniques are also highly relevant to inflammatory diseases of blood vessels. We have performed a series of early studies in Wegener's granulomatosis (WG). Using WCX-2 chips (Ciphergen Biosystems, Fremont, CA), we have analyzed 16 sera from eight WG patients, one sample from a period of active disease and another from remission for each patient. All patients had severe disease (defined as WG that constitutes an immediate threat to the patient's life or to vital organ function) at the time of initial sampling. Remission samples were obtained between 9 and 15 months after the start of treatment. The Birmingham Vasculitis activity scores for WG (19) for the patients had a mean of 8 (range: 4-14) during active disease, and was zero for all patients during remission.

The chip preparation protocol used (for WCX2 chips; Ciphergen Biosystems; Fremont, CA.) was designed to examine low MW proteins, particularly those with m/z ratios of less than 15,000. Figure 8 shows the serum proteomic profiles from baseline and remission in one of the patients, scanning all proteins with *m/z* ratios between 1,000 and 20,000. Even on visual inspection, major differences between these two profiles are evident at this magnification. First, in the baseline sample there is a narrow band of intensity around the *m/z* value 7,000 that is absent at remission. Second, the two peaks just below 15,000 are substantially more intense at baseline than in the remission sample. Third, in the remission sample, there is a broad range of intensity in the 2,500 – 3,000 range that is absent in the baseline sample. The contrasts in the proteomic profiles between states of active disease and remission, apparent even to visual inspection at this low power, are even more striking in magnified views (Fig. 9). In all 8 patients there was the consistent emergence of a peak in the region of MW 10,500 Da as the patient's clinical status changed from active disease to remission. The range of MWs tested in these protein chip assays represents only a small portion of the entire serum protein spectrum. Other MW ranges can be evaluated by slight alterations in the chip preparation techniques. Although these findings are preliminary, they underscore the potential of this technology when applied to well-characterized patient groups.

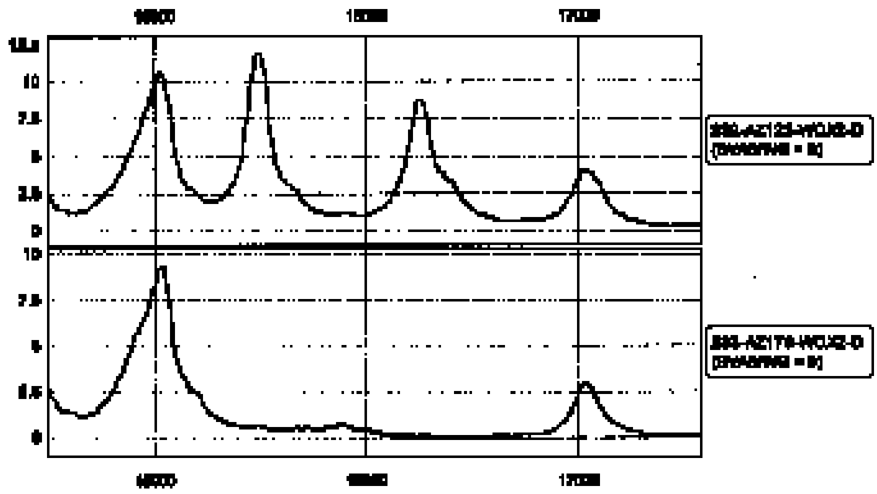

Fig. 9. A "magnified" view of the *m/z* ratios between 15,000 and 17,000 in the baseline and remission serum proteomic profiles from patient D. Two unequivocal peaks evident in the baseline sample, near 15,500 and 16,200. These peaks are completely absent in the sample from remission.

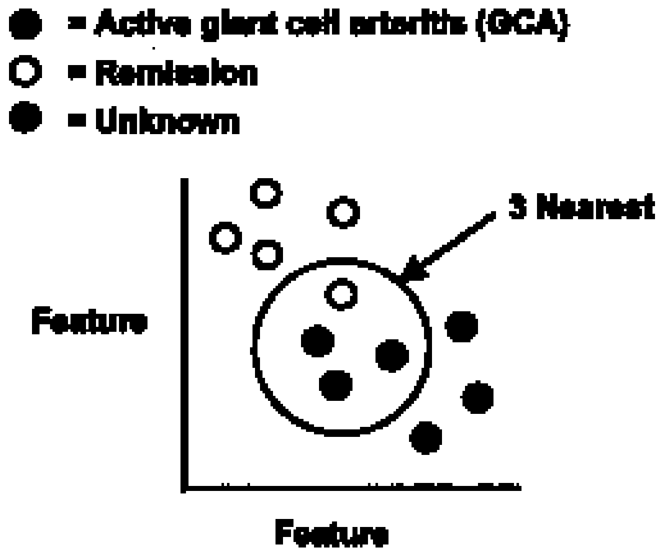

Fig. 10. Conceptual drawing of the concept of a clustering algorithm.

Bioinformatics

A variety of computational tools have been designed or adapted to mine the large amounts of data generated in proteomic analyses. Detailed discussions of these techniques are beyond the scope of this review (and, frankly, beyond the interest of most clinical investigators). What follows is an overview of the most common tools for analyzing proteomic data, with an emphasis on approaches to detecting patterns that segregate one state from another.

The sheer mountain of data points acquired by MS techniques can be overwhelming in size, complexity, dimensionality (i.e., number of data points), and computational requirement. A typical data file from a single sample generated from a “low resolution” MS technique (e.g., SELDI) has approximately 40,000 data points and a size of 800 KB. By way of comparison, a typical e-mail message is approximately 8 KB. Thus, the data contained within the file on one sample is equivalent to that contained within 100 e-mail messages (complex routing information and all). Even more daunting are files generated by high resolution instruments (e.g., tandem MS), which have approximately 350,000 data points and sizes of 5 MB for each sample. Each data point represents one *m/z* value and its corresponding intensity. The challenge lies in trying to identify the feature or features that differentiate one state from another. With very small sample sizes (e.g., n <30) it may be possible to inspect the samples visually and discern differences in peak intensities. As the example below shows, however, this approach is not practical for large numbers of samples.

Suppose that one is attempting to identify a combination of 5 *m/z* values to segregate two disease states (e.g., active Takayasu’s arteritis versus remission) using a low resolution mass spectrometer. In addition, assume that every possible combination of analytes will be analyzed, and that (hopefully) one has access to the world’s fastest supercomputer, which can perform 40 trillion calculations/second. Under such conditions, completion of the analysis would require nearly 9 months! Moreover, with data generated by a high resolution mass spectrometer and a combination of 10 *m/z* values designated to segregate the same two states, the analysis would take 6 x 10^{27} years. Clearly, the brute force method is not practical for such analyses. The field of bioinformatics is charged with parsing solutions to these challenges.

Specific bioinformatic approaches: Focus on clustering

Several computational strategies have been employed in proteomics analyses to date. In our own work, we have used an approach called "clustering". We discuss this strategy in some detail below. Other strategies employed include decision tree analysis, support vector machines, principal component analysis, and neural networks.

This practical example helps illustrate the clustering approach. Suppose that one would like to detect a proteomic fingerprint that segregates active giant cell arteritis (GCA) from remission and that one has a total of 246 samples (90 remission samples and 156 samples from patients with active GCA). The samples will be run on a high resolution mass spectrometer. One therefore anticipates approximately 86 million data points (246 samples x 350,000 data points/sample), comprising a potential total quantity of data of data of 1.2 GB. In order to produce a validated model, the samples must be divided into a training set and a testing set. As the names imply, the training data are used to build the model, and the testing data to validate it. The optimal model would have 100% sensitivity and specificity when applied to the testing data. For this example, we randomly divide the samples in half: 45 remission and 78 active GCA samples for both the training and testing phases. Figure 10 illustrates the concept of clustering of patient samples in multidimensional space according to the number of features examined (i.e., the specific number of ions (*m/z* values) used to discriminate clinical subsets).

We would like to detect a model that categorizes all remission samples and all active GCA samples into their own distinct groups. Figure 10 shows an example of the K-nearest neighbor (Knn) method. This figure shows two features, A and B, that are used to segregate the two groups. These features represent any *m/z* value. The samples are then plotted according to their corresponding intensity value for that particular feature. As Figure 10 indicates, the active GCAsamples appear to cluster in the lower portion of the graph. This means that, in general, active GCA samples have an intensity that is higher for Feature B, but lower for Feature A compared to the remission samples. The power of clustering comes when an unknown sample is then put through the model, as show by the black dot. The “K” in Knn represents the number of samples used to predict an unknown sample. In this case K = 3, which means that the 3 nearest neighbors are used to classify the sample. Thus, the black dot would be classified as a sample from a GCA patient, because 2 of the 3 nearest neighbors are samples from GCApatients (20). If the algorithm is trained correctly, clustering can be a very powerful prediction tool because of its natural ability to generalize.

Sources of variability in proteomic studies – and critical quality control measures

During the past few years, fantastic claims have been made for the derivation of diagnostic tests through proteomics. Most of these claims have not or will not stand up under further scrutiny (testing in new populations of patients, etc.). Unfortunately, relatively few papers have focused on quality control issues in proteomics and on methods of qualifying samples for analysis in the first place. Rigorous quality control efforts are essential to every stage, from the collection of samples to operating the MS instrument to the statistical analysis of data. The prediction power of bioinformatics algorithms is directly related to the quality of the data going in. The potential sources of error in proteomic studies include (but are not limited to):

* *Flawed procedures for the collection of sera.* As discussed above, allowing samples to sit for too long before processing is the cardinal offense in this category.
* *Improper calibration of instruments*. Standard operating procedures must be developed for the calibration of all MS instruments, which are inherently finicky.
* *Faulty protein chips or faulty individual spots on chips*. In many cases, quality control within the industry that produces commercially available protein chips and other implements has been poor. Application of a calibration sample to one spot on each protein chip may help overcome this problem. The spectra derived from calibration samples can then be compared against custom models designed by individual laboratories.
* *Failure to use control samples.* Control samples should be run with every study. In the case of protein chips, at least one control sample should be placed randomly on a spot for each chip. The control sample can be used to track the process variability – from sample preparation to mass spectrometer acquisition.
* Failure to assign samples randomly to training or testing sets. Samples should be randomized to either the training or testing phases. Clustering algorithms and other means of parsing proteomics data are very good at finding any difference between groups of interest. Without randomization of samples, differences detected between two sets of samples may have little to do with biologic plausibility and more to do with systematic handling differences in the samples.

Bench to bedside collaborations

The variety of skills needed to conduct cutting edge translational research in proteomics today calls for collaboration among individuals with expertise in many disparate fields. Indeed, the collaborative nature of proteomics investigations is a paradigm for the manner in which much good science is conducted today. The most productive work will derive from the joint efforts of scientists familiar with the type of rigorous laboratory techniques required, computer scientists who can design new bioinformatics approaches for this field, and clinical investigators who know what questions are relevant to patient care. For this third group of investigators, a thorough understanding of the disease of interest, the ability to provide reliable data on well-characterized patient cohorts, and a sufficient understanding of the technical issues of proteomics are all essential to effective collaborations.

Acknowledgement

The authors thank Dr. E.F. Petricoin, III, of the NCI/FDA Clinical Proteomics Program for many helpful discussions.

References

1. OSLER W: An address on high pressure. *Brit Med J* 1912; 2: 1.
2. PETRICOIN EF III, ARDEKANI A, HITT BA *et al.*: A bioinformatics analytical method reveals proteomic signatures of ovarian cancer in serum. *Lancet* 2002; 359: 572-7.
3. PETRICOIN EF III, ORNSTEIN DK, PAWELETZ CP *et al*.: Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2002; 94: 1576-8.
4. PUTNAM FW: *The Plasma Proteins Structure, Function, and Genetic Control.* Academic Press, New York, 1975-87, pp. 1-55.
5. ADKINS JN, VARNUM SM, AUBERRY KJ *et al.*: Toward a human blood serum proteome: Analysis by multi-dimensional separation coupled with mass spectrometry. *Mol Cell Proteom* 2003; 1: 947-55.
6. GORG A, WEISS W: Analytical IPG-Dalt. *Methods Mol Biol* 1999; 112: 189-95.
7. GORG A, OBERMAIER C, BOGUTH G *et al.*: The current state of two-dimensional electrophoresis with immobilize pH gradients. *Electrophoresis* 2000; 21: 1037-53.
8. GYGI SP, CORTHALS GL, ZHANG Y *et al*.: Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci USA* 2000; 17: 9390-5.
9. RICHTER R, SCHULZ-KNAPPE P, SCHRADER M *et al*.: Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J Chromatogr Biomed Sci Appl* 1999; 726: 25-35.
10. LEUNG S-M, ZEYDA T, THATCHER B: A new and rapid method for phenotype screening using Seldi Proteinchip™ arrays demonstrated on serum from knockout and wild type mice. *Mol Biol Cell* 1998; 9 (Suppl.): 351a.
11. PAWELETZ CP, GILLESPIE JW, ORNSTEIN DK *et al*.: Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Res* 2000; 49: 34-42.
12. WILKENS MR, WILLIAMS KL, APPELRD, DF H (Eds.): *Proteome Research: New Frontiers in Functional Genomics*. New York, Springer-Verlag, 1997.
13. AEBERSOLD R, MANN M: Mass spectrometry-based proteomics. *Nature* 2003; 422: 198-207.
14. STEEN H, KUSTER B, FERNANDEZ M *et al*.: Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal Chem* 2001; 73: 1440-8.
15. BALDWIN MA, MEDZIHRADSZKYKF, LOCK CM *et al*.: Matrix-assisted laser desorption/ionization coupled with quadrupole/orthogonal acceleration time-of-flight mass spectrometry for protein discovery, identification and structural analysis. *Anal Chem* 2001; 73: 1707-20.
16. MANN M, WILM MS: Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994; 66: 4390-9.
17. CONRADS TP, ISSAQ HJ, VEENSTRA TD:

New tools for quantitative phosphoproteome analysis. *Biochem Biophys Res Commun* 2002; 290: 885-90.

18. ONG SE, BLAGOEV B, KRATCHMAROVA I *et al*.: Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002; 1: 376-86.
19. STONE JH, HOFFMAN GS, MERKEL PA *et al*.: The Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS for WG): A disease-specific vasculitis activity index. *Arthritis Rheum* 2001; 44: 912-20.
20. PUNCH WF *et al*.: Further research on feature selection and classification using genetic algorithms. *Proceedings of the International Conference on Genetic Algorithms 1993*, University of Illinois, pages 557-64.
21. ANDERSON NL, ANDERSON NG: The human plasma proteome. *Mol Cell Proteom* 2002; 1: 845-67.