Published in final edited form as:

Mutat Res. 2011 June 17; 722(2): 171–182. doi:10.1016/j.mrgentox.2010.06.016.

Recent Developments in Quantitative Proteomics

Christopher H. Becker and

Caprion Proteomics U.S., 1455 Adams Drive, Menlo Park, CA 94025

Marshall Bern

Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304

Abstract

Proteomics is the study of proteins on a large scale, encompassing the many interests scientists and physicians have in their expression and physical properties. Proteomics continues to be a rapidly expanding field, with a wealth of reports regularly appearing on technology enhancements and scientific studies using these new tools. This review focuses primarily on the quantitative aspect of protein expression and the associated computational machinery for making large-scale identifications of proteins and their post-translational modifications. The primary emphasis is on the combination of liquid chromatography-mass spectrometry (LC-MS) methods and associated tandem mass spectrometry (LC-MS/MS). Tandem mass spectrometry, or MS/MS, involves a second analysis within the instrument after a molecular dissociative event in order to obtain structural information including but not limited to sequence information. This review further focuses primarily on the study of *in vitro* digested proteins known as bottom-up or shotgun proteomics. A brief discussion of recent instrumental improvements precedes a discussion on affinity enrichment and depletion of proteins, followed by a review of the major approaches (label-free and isotope-labeling) to making protein expression measurements quantitative, especially in the context of profiling large numbers of proteins. Then a discussion follows on the various computational techniques used to identify peptides and proteins from LC-MS/MS data. This review article then includes a short discussion of LC-MS approaches to three-dimensional structure determination and concludes with a section on statistics and data mining for proteomics, including comments on properly powering clinical studies and avoiding over-fitting with large data sets.

Keywords

quantitative proteomics; LC-MS; liquid chromatography-mass spectrometry; protein identification; affinity enrichment; affinity depletion

1. Introduction and Scope

The purpose of this review article is to report recent developments in proteomics, relevant to both clinical and basic research. Proteomics is a diverse field, with diverse techniques, and scientists outside the field may find it confusing and difficult to approach the subject.

Proteomics can be applied to any of the areas of biology and medicine. In the context of

Contact info for corresponding author: Christopher H. Becker, Caprion Proteomics U.S., 1455 Adams Drive, Suite 2126, Menlo Park, CA 94025, cbecker@caprion.com phone 650-325-9200 fax 650-325-9206.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

mutagenesis, the technology can be used to look for the translation of genetic mutations into changes in amino acid sequence as well as the impact of genetic mutations on resultant biology such as over-expression of proteins or alteration in post-translational modifications such as phosphorylation or methylation. This review however, focuses on the technological developments in modern proteomics and less on specific studies using the technology. The authors have tried to organize topics and recent developments in the most accessible fashion as they are able, forgoing any pretense of completeness.

The study of proteins and protein chemistry, the predecessors if you will of proteomics, extend back well over a century [1]. Proteomics as a new field, however, is really just blossoming, with numerous scientific journals dedicated to the subject, including *Proteomics*, *Proteomics – Clinical Applications*, *Journal of Proteomics*, *Journal of Proteome Research*, and *Molecular and Cellular Proteomics*, as well as articles employing and reviewing proteomics methods appearing in many other journals. A recent PubMed search on “proteomics” yielded over 20,000 citations. Many monographs have been written on the subject, some including integration of proteomics with genomics and bioinformatics; a search at the bookseller Amazon.com showed dozens of well-crafted monographs on the subject.

With this as backdrop, the goal of the present review is necessarily modest. First, we will attempt to delineate major topics within the field. Some of those topics will be covered by little more than their mention and a few references to more focused review articles and notable manuscripts. Some of the topics will be covered in greater depth, but none in any way that could be considered exhaustive. We thus acknowledge the limitations of this review, with apologies for the many intentional and unintentional omissions.

In order to organize topics and necessarily limit the scope of this review, we will begin by dividing the field of proteomics into quantitative expression proteomics (protein concentrations, including concentrations of their modifications) with associated peptide and protein identification, investigations of three-dimensional (3D) structure by mass spectrometry, affinity purifications, and the study of protein-protein interactions -- acknowledging that there is some overlap between these categories. We will focus on quantitative proteomics and associated protein identification, with additional comments and references on instrumentation, 3D structure by mass spectrometry, affinity enrichment, and relevant statistics and data mining. Regarding the study of protein-protein interactions, we merely cite here some notable reports, including protein complexes and molecular machines by mass spectrometry [2–5] and molecular biology techniques such as yeast-two-hybrid experiments [6].

Within the area of quantitative proteomics, we can further divide the field between mass spectrometric-based measurements with liquid or gel separations, and measurements made by protein affinity arrays and multiplexed immunoassays.

Protein arrays and multiplexed immunoassays are becoming more widespread and are increasing in content and capabilities. Some notable articles for protein affinity arrays and multiplexed immunoassays include [7–12].

This review, however, focuses on mass spectrometric-based assays for quantitative proteomics, especially the combination of high performance liquid chromatography (HPLC) and mass spectrometry, known as LC-MS and associated tandem mass spectrometry for peptide and protein identification, LC-MS/MS.

Top-down and bottom-up methods

Mass spectrometric measurements can be grossly divided into two categories depending on whether the final measurements are performed on intact or digested proteins. Measurements on intact proteins, can further be categorized depending on whether sequence structure and/or post-translational determinations are investigated, so-called top-down measurements [13,14], or whether the accurate molecular weight or distribution of weights in a mixture is all that is desired. Top-down measurements are gaining in capability but in the realm of proteomics where one generally deals with complex mixtures, generally only a limited number of proteins have been investigated and they must be at high relative concentration and not too massive. Further, generally very expensive Fourier-transform mass spectrometer (FTMS) instruments are required.

In contrast, investigation of proteins with the aid of *in vitro* enzymatic digestion, known as bottom-up or shotgun proteomics [15], represents the great majority of activity in the field and thus is the primary scope of this review. Reducing the size of polypeptides used in analysis aids in the ability to both quantify and identify a large number of proteins, and the smaller mass facilitates high resolution and effective use of a wide variety of instruments. Note however, that separations, enrichments and depletions of intact proteins prior to digestion can be part of this bottom-up proteomic approach. Also of note is the increasing ability of bottom-up proteomics to handle larger enzymatic fragments using an infrequent cutting protease.

In what follows, proteomic information is organized by:

- Instrumentation
- Affinity enrichment and depletion,
- Quantification techniques,
- Peptide and protein identification,
- 3D structure by mass spectrometry,
- Statistics and Data Mining.

2. Instrumentation

This review does not focus on instrumentation but it is certainly noteworthy that instrumentation continues to improve in terms of sensitivity, mass accuracy and resolution, dynamic range, speed of analysis, improved ionization and dissociation techniques, and software capabilities. Similarly liquid chromatography continues to evolve.

Resolution in simple terms is the sharpness of a mass spectral peak and is defined as $R = M/\Delta M$ the mass divided by the width of the spectral peak in mass; R is a dimensionless parameter. The resolution of the most expensive Fourier-transform mass spectrometers and the related newer-design Orbitrap mass spectrometer [16,17] reach 100,000 or more with ppm or sub-ppm accuracy. High mass accuracy can be further improved by post-acquisition calibration utilizing knowledge of identities of many of the molecules in the samples [18]. Resolution is also improving on other instruments, particularly time-of-flight mass spectrometers[19], reaching resolutions previously only thought to be available to Fourier-transform mass spectrometers. These instruments are built as both single analyzers and tandem mass spectrometers.

For profiling studies for quantitative proteomics, biomarker discovery and general biological investigation such as the focus here, high resolution is extremely useful, especially resolution sufficient to readily resolve isotopes of peptides. However, subsequent targeted

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

multiplexed and quantitative protein concentration measurements sometimes used for verification and validation studies are often performed by the method known as multiple reaction monitoring or MRM [20–22]. In the MRM method, a number of peptides of interest are chosen in advance and characteristic fragments of them are generated by energetic collision (see below) and then monitored; this procedure is not unlike in nuclear physics where a nucleus undergoes a collision and characteristic pieces emerge. High resolution can be useful for MRM but a common approach is to use LC-MS/MS by a triple quadrupole mass spectrometer with nominal unit mass resolution, although a linear ion trap [22] or other type of tandem mass spectrometer can also perform the MRM assay. For the MRM assay, because a limited though significant number of analytes are monitored, and established fragmentation transitions are used, unit mass resolution is still an effective approach. Commonly for the MRM method, peptides are synthesized with sequences identical to those being studied except that they contain heavy stable isotopes shifting the mass; known amounts can then be spiked into the sample before analysis to provide a basis for absolute quantification and further quality assurance.

Ionization is performed principally by two methods: electrospray ionization (ESI) and matrix-assisted laser desorption and ionization (MALDI). For ESI, a small flow of the analyte in solution is sprayed into a low-pressure chamber under charging conditions with a desolvation gas such as nitrogen present; finer and finer droplets form during the evaporation until the bare analyte ion remains and is introduced into the mass spectrometer. For MALDI, the analyte of interest is mixed with an excess of a small organic compound in solution; the solution is allowed to dry and then a pulse of laser light irradiates the sample, releasing the charged analyte for mass analysis. These discoveries in the late 1980's, opened the door to high mass measurements and were the subjects of the 2002 Nobel prizes in chemistry. (See the recorded speeches from the winners John Fenn and Koichi Tanaka that year for further background, http://nobelprize.org/nobel_prizes/chemistry/laureates/2002/.) It is possible to couple MALDI with liquid chromatography via spotting of eluate drops, but in general ESI is a more common approach to couple “on-line” liquid chromatography with mass spectrometry for LC-MS. A variety of off-line chromatographies can of course be combined with both ionization methods.

More recent has been a new development in terms of dissociation or fragmentation for the purposes of peptide and protein sequence identifications and associated post-translational modifications (PTMs). For many years, collision-induced dissociation or CID (sometimes known as collision-assisted dissociation or CAD) has been the workhorse of tandem mass spectrometry, breaking the analyte into characteristic fragments in order to provide this structural information. The CID method is the backbone of the MRM method, breaking the backbone of the peptide in characteristic places. While very powerful, CID has difficulty providing sequence information in certain cases such as for glycosylated or phosphorylated peptides, or larger polypeptides in high charge states. A significant recent development has been the advent of electron-transfer dissociation (ETD) for ion-trap mass spectrometers [23,24] and electron capture dissociation (ECD) under high vacuum such as for FTMS instruments [25]. In these instances the addition of an electron under gaseous conditions destabilizes the ion leading to dissociation. While not a panacea, these new methods are generally complementary and offer significant new capabilities for peptide and protein identification.

Liquid chromatography is improving in terms of available chemistries and physical properties such as higher pressure pumping systems to service higher resolution column packings with smaller diameter particles [26,27]. Better separations of analytes from each other give rise to more complete analyses with less interferences and ambiguities. Improvements include faster separations to help efficiency and greater sensitivity by

producing eluted molecules in a narrower time window increasing the eluted temporal concentration for enhanced signal.

For more information also contact the major LC and MS vendors such as (alphabetically) Agilent, Bruker, Shimadzu, Thermo Scientific, and Waters.

3. Affinity Enrichment and Depletion

A variety of enrichment or depletion methods are widely employed in conjunction with proteomic measurements. These can be categorized based on whether the affinity reagent is an antibody or other protein, or whether the reagent is a simple chemical. Enrichment is an approach used to increase sensitivity by a combination of collecting the targeted proteins or peptides while eliminating unwanted material, often in great excess; removing unwanted material lowers spectral interferences and competition for the ionizing charge needed for mass spectrometry. Depletion is an approach to remove unwanted large abundance molecules while nonselectively letting everything else remain.

In either case, the question arises for shotgun (bottom-up) proteomics as to whether to capture intact proteins and then digest, or to capture peptides from digested proteins. In most cases, both options are possible. As a rule, experimenters favor capturing peptides from digested proteins, especially for targeted PTMs. First, a premium is placed on actually observing the target structure in question including its position in the protein, and examining only peptides with the target structure avoids the ambiguous situation of capturing an intact protein and not knowing where/if the modification actually occurred. Second, reaction kinetics and steric interactions are more favorable for digested proteins. Important exceptions to this are the cases of capturing or depleting specific targeted proteins, and ubiquitinated proteins, where intact proteins must be captured. Note that even when a sandwich ELISA is available for a protein, sometimes immuno-mass spectrometry is useful to probe PTMs.

Many reagents have been described in the literature and are available commercially. Two of the more common PTMs targeted are glycosylation and phosphorylation. Glycosylation analysis (focusing on the peptide or glycan or the glycosylated form itself) frequently uses lectins as affinity reagents, such as concanavalin A (ConA) or wheat germ agglutinin (WGA). Another common approach for glycosylated peptides is capture by an azide resin [28] although the sugar has to be enzymatically severed in order to perform the analysis, so glycan information is sacrificed.

For phosphorylation, broad capturing of phosphorylated tyrosine is generally approached using a variety of commercial antibodies offered, for example, by Millipore Corp. and Cell Signaling Technology, while capture of phosphorylated serine and threonine is frequently accomplished by TiO₂ and some form of immobilized metal affinity chromatography (IMAC) [29].

But there are a great many more affinity capture approaches. For example, there are antibodies which can be used to capture 3-nitro-tyrosine (a rabbit polyclonal antibody from Millipore (via purchase of the former Chemicon) and a goat polyclonal antibody from Academy Bio-medical Co.) and a mass spectrometry study on a human pituitary adenoma has been conducted using the antibody from Chemicon [30]. Essentially, any epitope or functional group one can raise an antibody to, or can design a chemical coupling to, such as active sites of enzymes[31], can be captured.

Another example of sub-proteome enrichment is based on protein-affinity capture of the ubiquitin proteome. The ubiquitin-association domain of the human protein hPlc2 exhibits

broad polyubiquitin specificity yet with sufficient affinity to pull-down polyubiquitinated proteins; this reagent can conveniently be coupled via an NHS-ester-containing agarose bead [32]. This enabled researchers to perform differential expression of the ubiquitin proteome in response to disease and to pharmacological intervention; see the example of Figure 1.

Depletion of targeted proteins usually centers on the most abundant proteins in a complex mixture in order to increase the dynamic range of measurement for the remaining proteins. The most notable example is the removal of albumin and other abundant proteins from human and animal plasma or serum. Commercial products have been developed for this important case, offered by Agilent Technologies and Sigma-Aldrich. The detection limit for serum or plasma, arguably the most difficult biological fluid from a dynamic range perspective, can be as low as 1 ng/mL if extensive immunodepletion is coupled with multidimensional chromatography.

4. Quantification Techniques

Quantitative proteomics arguably began with 1-dimensional and 2-dimensional gel electrophoresis. The ability to identify proteins derived from visualized stained spots by mass spectrometry, notably at first by so-called peptide fingerprinting with MALDI, greatly improved the power of this method (see following Section 5), and it has continued to evolve and grow in capability with a greater variety of general and functional-group-specific gel stains, gel preparation methods and image analysis software. Nevertheless, liquid-based methods have now exceeded 2D gels in their capability to track and identify more proteins. Still, 2D gels can find particular utility in the study of post-translational modification, showing global behavior.

But quantitative LC-MS profiling [34], and associated identification, is the focus of this review, and the development of quantification for it has an interesting story. As LC-MS of peptides began to develop in the 1990's, the absence of a quantification method to rival 2D gels was apparent. A method for quantification in proteomics by LC-MS remained a puzzle for a long time because peptide mixtures can be extremely complex, with each peptide having different and unknown relative sensitivity factors (RSFs), and because workers in the field were well aware of so-called matrix effects where a molecule's RSF can change depending on the presence or absence of other molecules. At the same time, the use of isotopic-modified molecules has a long history in mass spectrometry for quantification, dating back to the 1930's, generating much of our knowledge of biochemical pathways [35,36]. Then in the late 1990's researchers hit upon an isotopic labeling strategy for differential expression, comparing two-samples by labeling them separately with a molecular tag containing a heavy or light isotope, mixing the samples so that they go through exactly the same chromatography and mass spectrometry, and then comparing the ratios of the measured heavy and light labeled peptides; this was the method known as isotope coded affinity tags or ICAT [37]. This approach has been refined with a method called iTRAQ (isobaric tag for relative and absolute quantification) [38]. Isotopic labeling rapidly expanded with a wide range of chemistries being explored where reactants with peptides could be isotope-labeled with stable heavy or light isotopes. Many different research groups have developed all sorts of other types of isotope labeling chemistries[34,39–41] including the relatively simple ^{18}O method, where one sample is digested in H_2^{18}O while the other is digested in ordinary aqueous solution to differentially label at the carboxyl terminus.

Another labeling approach has been to feed a cell line or small organism with isotope-labeled amino acids. The “stable isotope labeling by amino acids in cell culture” or SILAC

method [42] can be applied to organisms as large as mice, although in such instance the expense can be considerable for just a few animals.

Not very long after isotopic labeling became popular and accepted as a quantification method, approaches described as *label-free* were proposed and developed. At first this was based on the normalized intensities of the peptide signals [43–47], and then another approach called “spectral counting” came forward based on the frequency of identifications [48,49]. These methods met much resistance in the community that had become enamored of isotopic labeling, but gradually label-free methods have become widely accepted and practiced as their validity has been repeatedly demonstrated. Key to overcoming matrix effects was the idea of maintaining them to a constant degree by the use of significant chromatographic separation along with comparing similar types of samples. For example a public 2007 study by The Association of Biomolecular Resource Facilities (ABRF) compared various isotopic labeling and label-free approaches along with 2D gels in analyzing a set of blinded mixtures, with the results clearly showing the label-free method as successful; see <http://www.abrf.org/index.cfm/group.show/Proteomics.34.htm>.

Intensity-based (also known as MS1 or MS-only) label-free quantification is the more powerful of the two label-free quantification methods, but requires careful and reproducible sample preparation techniques along with sophisticated software that some small laboratories are not currently equipped to handle. The intensity-based label-free approach does not require continually re-identifying (redundantly identifying) peptides in every sample under study because it decouples profiling from identification and subsequently links the profiling and identification data sets *in silico* via accurate *m/z* and retention time. However, the spectral counting label-free method is the simplest to practice and thus is often favored in small laboratories; it suffers from significant uncertainties when the number of identifications per protein is small, such as 5 or less, and might be better termed semiquantitative to qualitative under those cases. Unfortunately most proteins in a complex mixture fall into the category of limited identifications, with only a few peptides or just one peptide identified; this is problematic for the scope of the spectral counting approach, along with the fact that every sample must undergo extensive MS/MS analysis.

As far as reproducible sample preparation, this is not difficult in principle; it is largely a matter of paying attention to making methods routine with appropriate assay testing. The development and use of standard operating procedures (SOPs), training of personnel, and testing of results with quality control samples are key ingredients. With regard to the software required for intensity-based label-free quantification, more and more software packages are appearing.

When laboratories use the intensity-based label-free approach for the first time, it is recommended that a pooled complex mixture be used, along with spikes of different concentrations of exogenous proteins in the pooled sample, to test preparation, LC-MS analysis and software tracking performance. Regular use of interspersed quality control samples is also an important practice to help obtain validated results.

Intensity-based label-free quantification is currently capable of differentially quantifying ~20,000 molecular ions in a single one-hour LC-MS run (counting only the monoisotopic molecular ions). The method can be used in two-dimensional chromatography arrangements as well, where each fraction from a first stage chromatography can produce this number of tracked molecular ions. Experienced laboratories can obtain average or median coefficients of variation (CVs) well under 10% for technical replicates (re-analysis of the same sample by LC-MS), and typically less than 10% for analysis of different aliquots from a pooled sample (including independent chemical processing plus LC-MS analysis). Biological

variation will of course increase CVs in a study, but it is common even for studies involving human clinical samples to see average within-cohort CVs of 25% or less, such as reported recently in a diagnostic/prognostic study of brain cancer using cerebral spinal fluid [50]. A typical number of proteins profiled with a false-discovery rate of 1% globally (see the following section on protein identification) for a single one-hour LC-MS run is 200–800, depending substantially on the nature of the specimen and preparation methods. Due to the masking effect of very abundant proteins, a one-hour LC run on blood plasma will typically profile fewer proteins than a similar run on a tissue sample. Adding additional fractions in a two-dimensional LC front-end analysis will substantially add to the number of proteins tracked and quantified, although not in a simple linear way as a function of the number of first-state chromatographic fractions.

MRM, as a targeted quantitative tool for validation or general selected targeted analysis, was mentioned above under instrumentation. Isotopically labeled peptides are commonly used as internal standards to increase quantification, precision and quality assurance.

5. Peptide and Protein Identification

(i) Historical Perspective

As early as the 1960s, researchers demonstrated the feasibility of low-throughput peptide and protein sequencing by mass spectrometry [51]. Klaus Biemann's work seems especially prescient. The chemical methods were quite different from today's methods, as they relied upon derivitization techniques to volatilize peptides, followed by gas chromatography, electron ionization, and in-source fragmentation, but the informatics stands out as surprisingly modern. Biemann and collaborators developed computer algorithms for *de novo* sequencing [52] and spectrum library searching [53]. They even correlated fragmentation spectra to gene sequences [54], but they did not have enough sequence to build anything that we would call a database.

With the inventions of MALDI and electrospray ionization in the late 1980s, mass spectrometry of biomolecules became much easier. Researchers first used the new ionization methods for *de novo* sequencing. Protein databases accelerated in growth and scope with the Human Genome Project in the early 1990s and other sequencing projects, and John Yates and Jimmy Eng published and patented SEQUEST [55], the first database search tool, in 1994. Moore et al. introduced reversed databases as decoys [56], although the technique gained popularity a few years later, when promoted by Gygi and others [57].

(ii) Peptide mass fingerprinting

This identification method is closely coupled to 2D gel proteomics. It can be considered the first method developed that used database information. The idea is to digest the protein(s) in a gel spot with an enzyme that cuts after specific residues, and identify proteins based on the pattern of masses of the resulting intact peptides. For example, using the most popular digestive enzyme, trypsin, which cuts after R (arginine) and K (lysine), one would expect to observe the peptides MQIFVK⁺ at 765.43 Da, TLTGK⁺ at mass 519.31 Da, and so forth from MQIFVKTLTGKTI... (ubiquitin). Typically 5 – 8 peptide masses measured at 50 ppm accuracy on a MALDI-MS instrument are sufficient to identify a protein of ~50,000 Da, with a few more peptides required for higher-mass proteins [58]. Peptide mass fingerprinting can identify simple mixtures of proteins, say 1 – 10 proteins, but it cannot handle complex mixtures, which produce too many unresolvable peaks and peptide ambiguities. Software for peptide mass fingerprinting includes ProFound [59] and Protein Prospector [60].

(iii) Identification from fragmentation spectra

Tandem mass spectrometry (MS/MS) has largely replaced peptide mass fingerprinting, even for gel proteomics, because it is much more sensitive and specific, capable of identifying proteins from only one or two peptides. Again the proteins are digested to peptides, but tandem mass spectrometry measures not just the m/z of the intact peptide ions (the “precursor” or “parent” ions), but also the m/z of fragments of selected precursors, usually the 5 or 10 most abundant precursors from the precursor scan. Tandem mass spectrometry benefits from two advantages, one physical and the other informatic: isolation of a narrow window of precursor m/z reduces background noise, and production of fragment ions greatly increases the information available for identification. The most important fragment ions are singly charged prefixes and suffixes of the peptide sequence, b- and y-ions for CID fragmentation and c- and z-ions for ETD fragmentation. These ions are numbered consecutively, so that b3 and c3 each include the 3 amino acid residues on the amino (N-terminal) end of the precursor peptide, and y4 and z4 each include the 4 amino acid residues on the carboxyl (C-terminal) end. The difference between b/y and c/z cleavage is in which bond is broken: the CO-NH or NH-CHR bond along the peptide backbone.

(iv) *De novo* sequencing

With complete fragmentation, it is in principle possible to determine the peptide sequence from the spectrum. Complete CID fragmentation for a peptide means either a b- or y-ion, but preferably both, for each peptide bond, for example, the b3 – b7 ions and the y2 – y7 ions for an 8-residue peptide. Even with complete fragmentation, however, determining the peptide is not a simple matter of reading off the sequence from a ladder of peaks, due to the b/y ambiguity of “signal” peaks and the distracting presence of “noise” peaks, which in effect include doubly charged ions, internal fragments, and neutral losses, as well as true chemical noise. See Figure 2. There are a number of good *de novo* sequencing programs including PEAKS from Bioinformatics Solutions [61] and PepNovo [62] from University of California at San Diego. Fragmentation is rarely complete, however, so even the best *de novo* sequencing programs are less sensitive than database search programs. Typically, *de novo* sequencers find a useful partial sequence, say 6 or more residues correct, on 60 – 90% of the spectra that can be identified by database search [63].

Recent research has shown that the prospects for successful *de novo* sequencing improve markedly with multiple spectra of the same peptide, especially complementary spectra showing various charge or modification states or the various ions produced by CID and ECD [64] or ETD [65] fragmentation. Multiple spectra give multiple chances to observe each cleavage, and at least in the case of CID/ETD pairs, can largely resolve the problem of disambiguating b- and y-ion peaks, one of the primary tasks of *de novo* sequencing. (If there is a peak at $m/z = x$ in the CID spectrum and a peak at $x+17$ in the ETD spectrum of the same peptide, then x is likely to be a b-ion. A peak at x in the CID spectrum and $x-16$ in the ETD spectrum is indicative of a y-ion.) We are not aware of any readily available *de novo* sequencing tools yet that take advantage of multiple spectra.

(v) Database searching

The workhorse of peptide identification is database searching, as implemented in such programs as SEQUEST [55], Mascot [66], X!Tandem [67], Protein Prospector, Phenyx, Spectrum Mill, etc. (There are at least 10 database search programs in active use.) All these programs compare observed spectra to predicted spectra for candidate peptides from a protein database. Genomic sequencing and years of proteomic study have made databases nearly complete for many species. Candidate peptides for each spectrum are selected by precursor mass and cleavage specificity, for example, the user may choose to consider all

semityptic peptides within ± 2 Daltons of the nominal precursor mass for a linear ion trap instrument (low resolution).

The user also decides what fixed and variable modifications to consider. Fixed modifications such as carbamidomethylated or carboxymethylated cysteine are assumed to be complete; whereas, variable modifications such as oxidized methionine are optional or partial, and the search will try both modified and unmodified. Most proteomics samples contain a variety of *in vivo* and *in vitro* variable modifications, but due to the combinatorial explosion of possibilities, which leads to loss of accuracy and sensitivity [68], it is best to run search engines with only the 4 or 5 most prevalent and/or interesting modifications enabled. Users often do not know these modifications in advance, and finding them may involve a certain amount of trial and error. Programs such as Popitam [69], InsPecT [70,71], and ByOnic [72] offer “blind” modification searching, which can find unanticipated modifications at some cost in speed, sensitivity, and accuracy.

(vi) Sequence Tagging

It is possible to combine *de novo* sequencing and database search in various ways. For example, one can compute a *de novo* sequence and then do a sequence search—something like BLAST—against the protein database; this is the approach taken by PEAKS and SPIDER [73]. Another technique computes numerous short sequence “tags” and then consults an index into the protein database to find peptides containing at least one of these tags; this is the approach taken by GutenTag [74], InsPecT [70], and Paragon [75]. Sequence tagging can give great computational speed-ups at some cost in sensitivity.

(vii) Spectrum Libraries

Peptide fragmentation sites are predictable to a certain extent, because low-energy CID preferentially breaks peptide bonds and gives only a small amount of side-chain fragmentation, but the details of a spectrum, such as the relative intensities of peaks, are mostly unpredictable. For example, it would be quite hard to predict that y_{14}^{++} and its neutral losses would be the tallest peaks in Figure 2. Spectrum library search uses previously identified real spectra rather than predicted spectra for peptide identification. Library search is faster than database searching, because even a large spectrum library contains fewer peptides than a protein database. Currently, the Global Proteome Machine [76] and PeptideAtlas [77], the largest libraries, contains tens of thousands of identified spectra of human peptides, but in contrast a human protein database contains millions of peptides, even counting only the fully tryptic peptides, and billions of possible modified peptides. Assuming high library coverage of the proteins in the sample, library searching also tends to be more sensitive than database searching, due to the focused set of candidate peptides and the fact that observed and re-observed spectra, especially those from the same type of instrument, are generally more similar than observed and theoretical spectra.

(viii) Scoring Algorithms

A step common to *de novo* sequencing, database searching, and spectrum library searching compares an unidentified observed spectrum with a theoretical or observed template spectrum. The scoring is at the heart of the identification algorithm. Due to the importance of the scoring step, we describe the three most common database-search scoring algorithms in more detail.

SEQUEST computes numbers called XCorr and Delta, where XCorr is essentially the dot product of the observed and theoretical spectra, divided by a cross-correlation with a random mass shift. In other words, XCorr’s numerator is a sum $\sum I(s)I(t)$, where s and t are observed and theoretical peaks, matching within a user-defined tolerance, say 0.4 Daltons, and $I(s)$

and $I(t)$ are functions of the peak intensities. SEQUEST considers theoretical peaks for a-, b-, and y-ions, and water losses from b- and y-ions. For the $I(t)$ functions, SEQUEST uses fixed intensities for theoretical spectra (for example, unit intensity for all b-ions) and “local intensities” for observed spectra, where “local intensity” measures a peak’s height relative to other peaks of similar m/z . Delta is the percent drop from the top XCorr to the second-best XCorr. An XCorr of 3.0 and a Delta of 0.05 generally indicates a correct identification.

Mascot uses a similar dot product, but it uses unit intensities for both theoretical and observed peaks. (None of the three popular search engines includes sequence-based intensity prediction, not even such well-understood phenomena as the tendency to fragment before proline.) Mascot decides where to cut the list of observed peaks adaptively, so that one spectrum could be scored by its top 40 peaks and another by its top 60 peaks. Mascot also chooses which types of peaks to score on a per-spectrum basis; this helps the program adapt to peptides that give mostly b-ions or give strong water losses. Mascot uses a probabilistic model to normalize scores, so that it can rationally determine whether matching 13 out of 30 predicted peaks is more or less significant than matching 20 out of 50; it then converts probabilities back to scores ranging from about 10 to 100, with a score of 40 or more generally indicating a correct identification.

X!Tandem is fully open-source software. X!Tandem uses a raw peak-matching score called a hyperscore [78], and then maps these scores to E-values (the expected number of random matches to the spectrum with such a good score) by fitting a curve to all the peptide hyperscores and measuring how far out on the tail of this distribution is the top-scoring peptide. Thus in some sense X!Tandem’s E-value is “all Delta” and “no XCorr”, and automatically compensates for spectra with differing amounts of signal and noise. Due to undersampling of the score distribution, however, X!Tandem’s E-values tend to be unreliable on very small searches and on weak matches with E-values above 1.0; a potential fix to this problem uses *de novo* techniques to estimate the random-peptide score distribution for a given spectrum [79].

Better scoring algorithms are an area of active research, with some groups focusing on improved intensity prediction as the key to greater sensitivity/specificity. In our own work on ByOnic [72], however, we have found that relatively simple additions to the basic dot-product scorer yield just as great performance improvements as sophisticated or even “perfect” (spectrum library) intensity prediction. For example, we have found that incorporating a mass-error term into the dot product makes a substantial difference. In Mascot, SEQUEST, and X!Tandem, peaks matching within 0.1 Da score no better than peaks matching within 0.2 Da. Similarly, it is advantageous to recalibrate m/z measurements [63], both precursors and fragments, based on confident identifications from the same LC run. As shown in Figure 3, recalibration is especially effective for high resolution TOF (time-of-flight) and FTMS (Fourier transform mass spectrometry) measurements.

(ix) Significance Analysis

The latest versions of most of the major search engines output a statistical significance estimate (a p-value or E-value) as well as raw scores, but a better estimate and more accurate filtering of false positives can be achieved by considering the entire ensemble of scores and identifications. PeptideProphet [80], SILVER [81,82], and Percolator [83] are three tools that “rescore” peptide identifications, taking into account extra information, such as the score distribution of all the spectra, the precursor m/z errors of confident identifications, and so forth.

False positive and false discovery rates (FDRs) [84,85] on a given data set can be estimated empirically by searching a database containing “decoys”—false proteins or peptides. A

simple and powerful way to create a decoy database, statistically matched for total size, number of tryptic peptides, and amino acid frequencies, is to simply reverse each protein in the original database [56]. The decoy technique works at both the peptide and protein levels, so that one can send the decoys through a succession of tools (for example, database search, significance analysis, and protein assembly) in order to measure the false discovery rate of the complete pipeline.

(x) Protein Assembly

Except in peptidomics studies, the spectrum and peptide levels of detail are artificial, an artifact of the bottom-up approach to proteomics. The meaningful level is the protein level, so an important step in the informatics pipeline is integration of peptide identifications into protein identifications. Specialized tools for protein assembly include Qscore [56], DTASelect [86], ProteinProphet [87] (and its commercial descendant, Scaffold), and ComByne [88]. Protein assembly tools typically rank proteins by confidence, treating distinct peptides as independent evidence of the presence of a protein, but discounting duplicate identifications. These tools output protein groups, rather than single proteins, in the case that the peptide identifications do not distinguish between homologous proteins. A committee of experts drafted the “Paris guidelines” for presentation of proteomics results [89]. These guidelines recommend that all the proteins sharing the same spectral matches be shown in a group, and that a protein with a strict subset of these spectral matches be dropped from the group for lack of independent evidence.

6. Mass spectrometry for 3D structure

Some experimental techniques based on mass spectrometry can give structural information about biological molecules and complexes [5]. These techniques do not solve structures to the level of atomic coordinates as in x-ray crystallography or NMR, but rather they measure solvent accessibility or provide distance constraints. This modest amount of information may be all that is needed for insight into molecular function, especially when used for differential analysis, comparing the molecule in two different states [5].

Mass-spec-based structural techniques include hydrogen-deuterium (H-D) exchange and covalent labeling methods. H-D exchange relies on the fact that amide hydrogen atoms along the backbone of a protein exchange with hydrogen from water, and the rate of exchange correlates with solvent accessibility and secondary structure stability. If the solvent is deuterated (heavy) water rather than ordinary water, the swapped-in deuterium atoms can be detected and quantified by mass spectrometry. H-D exchange is carried out at neutral pH, and the exchange is quenched by dropping both the pH and the temperature to prevent back-exchange. Analysis usually involves pepsin digestion, followed by electrospray ionization and single-MS profiling, to obtain resolution at the peptide level. This analysis is preferred over the usual shotgun proteomics method of trypsin digestion followed by CID MS/MS, because trypsin does not perform well at low pH and CID fragmentation tends to scramble the deuterium. ETD MS/MS, however, can be used to give single-residue resolution [90].

In a covalent labeling method, the experimenter induces a chemical reaction that makes a permanent modification to the protein that is indicative of structure. For example, the experimenter might acetylate all the exposed lysines, and then assay which lysines were exposed using standard shotgun proteomics [91]. “Oxidative footprinting” [92] uses hydroxyl radicals, produced by splitting water or hydrogen peroxide, to modify (oxidize) the protein. This method yields many more structural probes, because at least 6 of the 20 amino acid residues oxidize readily under hydroxyl radical attack.

Finally, cross-linkers offer a way to determine distance constraints: which amino acid residues are close to each other in the solution structure(s). Cross-linkers can be made in various lengths and with various residue specificities. The challenge in cross-linker experiments is to produce enough cross-linked peptides to provide a significant amount of structural information [5].

7. Statistics and Data Mining

The data sets from proteomics experiments are often large, with mass spectra from one sample's LC-MS analysis occupying a gigabyte or more of storage. After a data set has been processed, it can be reduced to a compact table with intensity information of specific peptides (and their PTMs) along with their associated identifications (if known), mass-to-charge ratios (m/z), charge states (z), and chromatographic retention times or characteristics. At this point, data will be ready for various statistical analyses. There might be only a few measurements per sample in an MRM experiment, but in proteomic profiling experiments there are often many thousands of individual peptide measurements per sample. In any event, it is useful to divide consideration of the statistics into univariate and multivariate approaches.

Often analyses use $\log_2(\text{intensity})$ as the dependent variable. The log transformation is used when the observed distribution of intensity data is closer to lognormal than normal. Data can be tested for normality by techniques such as the Shapiro-Wilk test [93].

(i) Univariate Statistics

Analyses of variance/covariance (AN(C)OVAs) are typically performed with all appropriate cohort categories as covariates. For simplicity we refer to all of these calculations as ANOVA methods even in cases where ANCOVA would be customary. Models should include a block effect when a randomized block experimental run order design [94] is employed—a highly recommended approach. (Failure to use a block randomization approach has been implicated in early poor results for clinical proteomics [95].) The p-values are computed from these ANOVA methods, with the results reflecting the evaluation of the hypothesis that the tested effect is zero, with other effects held constant.

Other statistics include fold-change, often computed as the median of cohort intensity ratios, effect size, and area under the receiver-operator characteristic curve (ROC_AUC). The ROC_AUC is a single number that facilitates comparisons between the sensitivity and specificity for a given variable. An area of 1.0 means complete separation between cohorts, while an area of 0.5 signifies no separation. The ROC_AUC can be conveniently computed as $\max(U/(n_1*n_2), 1-U/(n_1*n_2))$, where U is the Mann-Whitney statistic, equal to the number of pairs $(x[i], y[j])$ where $y[j] \leq x[i]$ [96,97].

A common occurrence is when one peptide molecular ion (sometimes called a spectral component) is observed in one cohort but not another. In such cases a p-value can still be generated by testing the hypothesis that observed missing values for a given component (if any) are distributed randomly between the two groups, using Fisher's exact test. Low p-values for this calculation indicate a greater degree of segregation of missing values into just one group than would be likely under pure randomization.

For situations with only small numbers of samples per cohort, such as common with *in vitro* experiments, more stable results can sometimes be obtained from methods that bring in information from the entire set of components when estimating standard deviations and t-ratios (sometimes referred to as “borrowing strength” across components). The moderated t-

statistic of Smyth [98] is a well known example of such a method. The standard deviations obtained from such methods have resistance to outliers.

Here and in other contexts, we face the issue of adjusting p-values to account for multiple comparisons. Statistical procedures to control the so-called False Discovery Rate (FDR) aim to maximize the number of true positive statistical results while controlling the proportion of false positives; see for example [99–103]. FDR procedures have become well known in the analysis of microarray data and are appropriately applied to proteomic data. Significance analysis of microarrays (SAM) can be repurposed for proteomic differential expression studies. SAM uses a slightly modified t-test, and computes permutations of replicate measurements (in silico sampling) to estimate FDR [104]. (The FDR for statistical analysis of group changes is distinct from the FDR used in peptide/protein identification.) FDR computations tend to err on the conservative side with shotgun proteomic data when multiple components are detected for a given protein unless a model is used to evaluate the effect of multiple consistent measures.

(ii) Multivariate Statistics and Data Mining

The analysis of data sets consisting of thousands of measurements made on a much smaller number of samples presents numerous challenges. Experimental results of this kind are sometimes referred to as “large p, small n” data; in the proteomic case, p would be the number of molecular ions quantified, and n the number of samples analyzed. Large p, small n results arise in several contexts in addition to proteomics, such as microarray studies, image analysis, and high throughput molecular screening. It is important to take advantage of methods that address the issues arising from the very high dimensional nature of such data. The number of appropriate available multivariate techniques is quite large, but most of those we find useful can be viewed as performing variable selection (or, more generally, dimension reduction), unsupervised learning or clustering, and supervised learning or discriminant analysis. These categories are not mutually exclusive, because some techniques can be viewed as performing functions in more than one category.

A variety of clustering and discriminant analysis techniques are useful, such as support vector machines (SVMs) [105], Breiman’s Random Forest [106] and Tibshirani’s prediction analysis for microarrays (PAM) [107]. With all the discriminant analysis methods, it is important to guard against the danger of overfitting, that is, producing a (typically over-complex) model that yields near-perfect classifications by sheer coincidence in the particular data set being analyzed, but that is very unlikely to work when applied to new data. A widely used precaution against overfitting is cross-validation. Fitting algorithms are then evaluated based on their cross-validated prediction accuracy. Many recent classification methods have cross-validation steps built in; for methods that do not, it is important to use cross-validation in some form to perform model selection. The ultimate validation approach of course is to perform additional measurements on a distinct, and preferably larger, population or sample set.

A generalized computational workflow for a quantitative proteomic pipeline with a range of approaches is shown in Figure 4.

Note several recent methods can integrate biological information such as Gene Ontology (GO) annotations or KEGG pathways into the analysis. One example [108]; is a method called GSCGRDA (group-shrinkage shrunken centroids group regularized discriminant analysis), which shares ideas with PAM but uses pathway information to obtain a better estimate of the correlation structure between genes, and appears to offer the potential for superior performance in cancer classification.

(iii) Requirement of Cohort Size

Especially for clinical and pre-clinical specimens, when designing a study, consideration must be made to the choice of cohort size, that is, the number of samples to be used for each cohort. However, one is often constrained by specimen and/or budget availability. We consider requirements for needed cohort sizes now.

The effect size is defined as $E = (\mu_Y - \mu_X)/\sigma$, i.e., the difference between population means measured in units of the population standard deviation. A power calculation is made for the study, and commonly a significance level of 0.01 or 0.05 is chosen along with a customary requirement of achieving 80% power (power can be defined as the probability that a statistical significance test will reject the null hypothesis for a specified value of an alternative hypothesis). As examples, results for power calculations for $n = 15$ and 30 are provided in Figure 5 for the case of paired comparisons. The calculations result in a detectable effect size of about 1.3 for $n = 15$. For the larger cohort size of 30, the detectable effect size is about 0.9. Thus, if the coefficient of variation for the measurement (including biological and analytical variations) is 30% and the detectable effect size is 1.3, then a minimum detectable fold-change is $1.3 = \text{fold-change}/0.3$ or a minimum fold change of 0.39.

8. Conclusions

Proteomics has undergone a rapid growth and maturation, following in the footsteps of genomics. Scientists demand quantitative results and the field of proteomics has responded with developments for differential expression (profiling) and large-scale peptide and protein identification. Intense effort is still being applied throughout the field with rapid improvements in areas such as the analysis of post-translational modifications using affinity capture techniques and advanced tandem mass spectrometry methods, as well as instrumentation and three-dimensional structure determination. Statistical and data mining methods familiar in other scientific disciplines have also found their way into the mainstream of proteomics.

Proteins are the main actors in the world of biology and these developments in quantitative proteomics bode well for a continued rapid expansion of our knowledge in fields critical to improving human health and other biology-related fields.

Acknowledgments

The authors are pleased to acknowledge the many contributions of Drs. Ted Jones, David Goldberg, Tom Shaler, Hua Lin, Mimi Roy, Shanhua Lin, Haihong Zhou, Weixun Wang, Howard Schulman, Eric Bennett and Ron Kopito. Christopher Becker acknowledges his research was supported in part with Federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200800059C. Marshal Bern acknowledges his research was supported in part by NIH grant GM085718.

References

1. Tanford, C.; Reynolds, JA. *Nature's robots : a history of proteins*. Oxford University Press; Oxford ; New York: 2001.
2. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. *Nature*. 2008; 453:1262–1265. [PubMed: 18563089]
3. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprpto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP. The molecular architecture of the nuclear pore complex. *Nature*. 2007; 450:695–701. [PubMed: 18046406]
4. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM,

- Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 2006; 440:631–636. [PubMed: 16429126]
5. Chance, M. Mass spectrometry analysis for protein-protein interactions and dynamics. Wiley; Hoboken, N.J.: 2008.
6. Brückner APC, Lentze N, Auerbach D, Schlattner U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int J Mol Sci*. 2009; 10:2763–2788. [PubMed: 19582228]
7. Ramachandran N, Raphael JV, Hainsworth E, Demirkan G, Fuentes MG, Rolfs A, Hu Y, LaBaer J. Next-generation high-density self-assembling functional protein arrays. *Nat Methods*. 2008; 5:535–538. [PubMed: 18469824]
8. Mulder J, Bjorling E, Jonasson K, Wernerus H, Hober S, Hokfelt T, Uhlen M. Tissue profiling of the mammalian central nervous system using human antibody-based proteomics. *Mol Cell Proteomics*. 2009
9. Fasolo J, Snyder M. Protein microarrays. *Methods Mol Biol*. 2009; 548:209–222. [PubMed: 19521827]
10. Patwa TH, Qiu Y, Zhao J, Simeone DM, Lubman DM. All-liquid separations, protein microarrays, and mass spectrometry to interrogate serum proteomes: an application to serum glycoproteomics. *Methods Mol Biol*. 2009; 520:75–87. [PubMed: 19381948]
11. Hu J, He X, Baggerly KA, Coombes KR, Hennessy BT, Mills GB. Non-parametric quantification of protein lysate arrays. *Bioinformatics*. 2007; 23:1986–1994. [PubMed: 17599930]
12. Gu Y, Zeleniuch-Jacquotte A, Linkov F, Koenig KL, Liu M, Velikokhatnaya L, Shore RE, Marrangoni A, Toniolo P, Lokshin AE, Arslan AA. Reproducibility of serum cytokines and growth factors. *Cytokine*. 2009; 45:44–49. [PubMed: 19058974]
13. Boyne MT, Garcia BA, Li M, Zamdborg L, Wenger CD, Babai S, Kelleher NL. Tandem mass spectrometry with ultrahigh mass accuracy clarifies peptide identification by database retrieval. *J Proteome Res*. 2009; 8:374–379. [PubMed: 19053528]
14. Breuker K, Jin M, Han X, Jiang H, McLafferty FW. Top-down identification and characterization of biomolecules by mass spectrometry. *J Am Soc Mass Spectrom*. 2008; 19:1045–1053. [PubMed: 18571936]
15. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR 3rd. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem*. 1997; 69:767–776. [PubMed: 9043199]
16. Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev*. 2008; 27:661–699. [PubMed: 18683895]
17. Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem*. 2000; 72:1156–1162. [PubMed: 10740853]
18. Becker CH, Kumar P, Jones T, Lin H. Nonparametric mass calibration using hundreds of internal calibrants. *Anal Chem*. 2007; 79:1702–1707. [PubMed: 17297976]
19. Vestal ML. Modern MALDI time-of-flight mass spectrometry. *J Mass Spectrom*. 2009; 44:303–317. [PubMed: 19142962]
20. Anderson L, Hunter CL. Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Mol Cell Proteomics*. 2006; 5:573–588. [PubMed: 16332733]
21. Barr JR, Maggio VL, Patterson DG Jr, Cooper GR, Henderson LO, Turner WE, Smith SJ, Hannon WH, Needham LL, Sampson EJ. Isotope dilution--mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I. *Clin Chem*. 1996; 42:1676–1682. [PubMed: 8855153]
22. Lin S, Shaler TA, Becker CH. Quantification of intermediate-abundance proteins in serum by multiple reaction monitoring mass spectrometry in a single-quadrupole ion trap. *Anal Chem*. 2006; 78:5762–5767. [PubMed: 16906721]
23. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*. 2004; 101:9528–9533. [PubMed: 15210983]

24. Pitteri SJ, Chrisman PA, Hogan JM, McLuckey SA. Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: reactions of doubly and triply protonated peptides with SO₂⁺. *Anal Chem.* 2005; 77:1831–1839. [PubMed: 15762593]
25. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem.* 2000; 72:563–573. [PubMed: 10695143]
26. Plumb R, Mazzeo JR, Grumbach ES, Rainville P, Jones M, Wheat T, Neue UD, Smith B, Johnson KA. The application of small porous particles, high temperatures, and high pressures to generate very high resolution LC and LC/MS separations. *J Sep Sci.* 2007; 30:1158–1166. [PubMed: 17595951]
27. Jemal M, Xia YQ. LC-MS Development strategies for quantitative bioanalysis. *Curr Drug Metab.* 2006; 7:491–502. [PubMed: 16787158]
28. Zhang H, Li XJ, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol.* 2003; 21:660–666. [PubMed: 12754519]
29. Bodenmiller B, Mueller LN, Mueller M, Domon B, Aebersold R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Methods.* 2007; 4:231–237. [PubMed: 17293869]
30. Zhan X, Desiderio DM. Nitroproteins from a human pituitary adenoma tissue discovered with a nitrotyrosine affinity column and tandem mass spectrometry. *Anal Biochem.* 2006; 354:279–289. [PubMed: 16777052]
31. Cravatt BF, Wright AT, Kozarich JW. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu Rev Biochem.* 2008; 77:383–414. [PubMed: 18366325]
32. Bennett EJ, Shaler TA, Woodman B, Ryu KY, Zaitseva TS, Becker CH, Bates GP, Schulman H, Kopito RR. Global changes to the ubiquitin system in Huntington's disease. *Nature.* 2007; 448:704–708. [PubMed: 17687326]
33. Bennett EJ, Shaler TA, Becker CH, Schulman H, Kopito RR. Unpublished results. 2007
34. Sechi, S. Quantitative proteomics by mass spectrometry. Humana Press; Totowa, N.J.: 2007. p. xp. 218
35. Hevesy G, Hofer E. Elimination of Water from the Human Body. *Nature.* 1934; 134:879.
36. Caprioli RM. Use of Stable Isotopes. *Biochem Appl Mass Spectrom.* 1972:735–776.
37. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol.* 1999; 17:994–999. [PubMed: 10504701]
38. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlett-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics.* 2004; 3:1154–1169. [PubMed: 15385600]
39. Julka S, Regnier F. Quantification in proteomics through stable isotope coding: a review. *J Proteome Res.* 2004; 3:350–363. [PubMed: 15253416]
40. Julka S, Regnier FE. Recent advancements in differential proteomics based on stable isotope coding. *Brief Funct Genomic Proteomic.* 2005; 4:158–177. [PubMed: 16102271]
41. Ye X, Luke B, Andresson T, Blonder J. 18O stable isotope labeling in MS-based proteomics. *Brief Funct Genomic Proteomic.* 2009; 8:136–144. [PubMed: 19151093]
42. Oda Y, Nagasu T, Chait BT. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat Biotechnol.* 2001; 19:379–382. [PubMed: 11283599]
43. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem.* 2003; 75:4818–4826. [PubMed: 14674459]
44. Wiener MC, Sachs JR, Deyanova EG, Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal Chem.* 2004; 76:6085–6096. [PubMed: 15481957]
45. Roy SM, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling. *Methods Mol Biol.* 2007; 359:87–105. [PubMed: 17484112]

46. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*. 2004; 3:984–997. [PubMed: 15269249]
47. Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*. 2005; 6:179. [PubMed: 16026613]
48. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics*. 2005; 4:1487–1502. [PubMed: 15979981]
49. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*. 2004; 76:4193–4201. [PubMed: 15253663]
50. Roy S, Josephson SA, Fridlyand J, Karch J, Kadoch C, Karrim J, Damon L, Treseler P, Kunwar S, Shuman MA, Jones T, Becker CH, Schulman H, Rubenstein JL. Protein biomarker identification in the CSF of patients with CNS lymphoma. *J Clin Oncol*. 2008; 26:96–105. [PubMed: 18056677]
51. Kinter, MaSNE. Protein Sequencing and Identification Using Tandem Mass Spectrometry. John Wiley & Sons; 2000.
52. Biemann K, Cone C, Webster BR, Arsenault GP. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J Am Chem Soc*. 1966; 88:5598–5606. [PubMed: 5980176]
53. Hertz HS, Hites RA, Biemann K. Identification of mass spectra by computer-searching a file of known spectra. *Analytical Chemistry*. 1971; 43:681–691.
54. Herlihy WC, Royal NJ, Biemann K, Putney SD, Schimmel PR. Mass spectra of partial protein hydrolysates as a multiple phase check for long polypeptides deduced from DNA sequences: NH₂-terminal segment of alanine tRNA synthetase. *Proc Natl Acad Sci U S A*. 1980; 77:6531–6535. [PubMed: 7005898]
55. Eng J, McCormack AL, Yates JR. An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994; 5:976–989.
56. Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom*. 2002; 13:378–386. [PubMed: 11951976]
57. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007; 4:207–214. [PubMed: 17327847]
58. Siuzdak, G. The Expanding Role of Mass Spectrometry in Biotechnology. MCC Press; San Diego: 2003.
59. Zhang W, Chait BT. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem*. 2000; 72:2482–2489. [PubMed: 10857624]
60. Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*. 1999; 71:2871–2882. [PubMed: 10424174]
61. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*. 2003; 17:2337–2342. [PubMed: 14558135]
62. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem*. 2005; 77:964–973. [PubMed: 15858974]
63. Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comput Biol*. 2006; 13:364–378. [PubMed: 16597246]
64. Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. Proteomics-grade de novo sequencing approach. *J Proteome Res*. 2005; 4:2348–2354. [PubMed: 16335984]
65. Datta, RaBM. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *Journal of Computational Biology*. 2009; 16
66. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20:3551–3567. [PubMed: 10612281]
67. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004; 20:1466–1467. [PubMed: 14976030]

68. Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*. 2004; 4:619–628. [PubMed: 14997485]
69. Hernandez P, Gras R, Frey J, Appel RD. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*. 2003; 3:870–878. [PubMed: 12833510]
70. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*. 2005; 77:4626–4639. [PubMed: 16013882]
71. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*. 2005; 23:1562–1567. [PubMed: 16311586]
72. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal Chem*. 2007; 79:1393–1400. [PubMed: 17243770]
73. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J Bioinform Comput Biol*. 2005; 3:697–716. [PubMed: 16108090]
74. Tabb DL, Saraf A, Yates JR 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem*. 2003; 75:6415–6421. [PubMed: 14640709]
75. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics*. 2007; 6:1638–1655. [PubMed: 17533153]
76. Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom*. 2005; 19:1844–1850. [PubMed: 15945033]
77. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. *Nucleic Acids Res*. 2006; 34:D655–658. [PubMed: 16381952]
78. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*. 2003; 17:2310–2316. [PubMed: 14558131]
79. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*. 2008; 7:3354–3363. [PubMed: 18597511]
80. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002; 74:5383–5392. [PubMed: 12403597]
81. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol*. 2004; 22:214–219. [PubMed: 14730315]
82. Gibbons FD, Elias JE, Gygi SP, Roth FP. SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *J Am Soc Mass Spectrom*. 2004; 15:910–912. [PubMed: 15144981]
83. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007; 4:923–925. [PubMed: 17952086]
84. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7:29–34. [PubMed: 18067246]
85. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*. 2008; 7:40–44. [PubMed: 18052118]
86. Tabb DL, McDonald WH, Yates JR 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res*. 2002; 1:21–26. [PubMed: 12643522]
87. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–4658. [PubMed: 14632076]

88. Bern M, Goldberg D. Improved ranking functions for protein and modification-site identifications. *J Comput Biol.* 2008; 15:705–719. [PubMed: 18651800]
89. Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics.* 2006; 5:787–788. [PubMed: 16670253]
90. Zehl M, Rand KD, Jensen ON, Jorgensen TJ. Electron transfer dissociation facilitates the measurement of deuterium incorporation into selectively labeled peptides with single residue resolution. *J Am Chem Soc.* 2008; 130:17453–17459. [PubMed: 19035774]
91. Scholten A, Visser NF, van den Heuvel RH, Heck AJ. Analysis of protein-protein interaction surfaces using a combination of efficient lysine acetylation and nanoLC-MALDI-MS/MS applied to the E9:Im9 bacteriotoxin--immunity protein complex. *J Am Soc Mass Spectrom.* 2006; 17:983–994. [PubMed: 16713291]
92. Xu G, Chance MR. Hydroxyl radical-mediated modification of proteins as probes for structural proteomics. *Chem Rev.* 2007; 107:3514–3543. [PubMed: 17683160]
93. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* 1965; 52:591–611.
94. Hinkelmann, K.; Kempthorne, O. *Design and Analysis of Experiments: Introduction to experimental design.* John Wiley and Sons; 1994.
95. Baggerly KA, Coombes KR, Neeley ES. Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J Clin Oncol.* 2008; 26:1186–1187. author reply 1187–1188. [PubMed: 18309960]
96. Hollander, M.; Wolfe, DA. *Nonparametric statistical methods.* Wiley; 1973.
97. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988:837–845. [PubMed: 3203132]
98. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004; 3 Article3.
99. Benjamini J, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Royal Stat Soc B Met.* 1995; 57:289–300.
100. Grant GR, Liu J, Stoekert CJ Jr. A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics.* 2005; 21:2684–2690. [PubMed: 15797908]
101. Qian HR, Huang S. Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics.* 2005; 86:495–503. [PubMed: 16054333]
102. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics.* 2003; 31:2013–2035.
103. Yang JJ, Yang MC. An improved procedure for gene selection from microarray experiments using false discovery rate criterion. *BMC Bioinformatics.* 2006; 7:15. [PubMed: 16405735]
104. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001; 98:5116–5121. [PubMed: 11309499]
105. Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines.* Cambridge University Press; 2000.
106. Breiman L. Random Forests. *Machine Learning.* 2001; 45:5–32.
107. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A.* 2002; 99:6567–6572. [PubMed: 12011421]
108. Tai F, Pan W. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics.* 2007; 23:3170–3177. [PubMed: 17933851]

Figure 1.

Results from cells in culture treated with a proteasome inhibitor (ALLn, N-Acetyl-Leu-Leu-Nle-CHO, 25 micrograms/mL) with the polyubiquitinated proteome enriched by affinity capture with differential expression quantification performed. A system biology analysis of several hundred polyubiquitinated proteins shows a shift in this proteome from protein synthesis and other functions to protein degradation[33]. Functional categories: **0hr pre-dose**, A protein synthesis 48%, B RNA metabolism/transport 16%, C sterol metabolism 2%, D development 2%, E transcription 2%, F ion homeostasis 2%, G cell respiration 2%, H protein folding 3%, I redox homeostasis 2%, J DNA replication 2%, K cell signaling 7%, L cytoskeleton 7%, M protein catabolism 3%, N apoptosis 2%; **1hr post-dose**, A protein catabolism 43%, B apoptosis 2%, C protein synthesis 21%, D RNA metabolism/transport, E chromatin structure 2%, F unknown 8%, G sterol metabolism 2%, H transcription 5%, I protein folding 2%, J cell signaling 2%, K cytoskeleton 5%; **8hr post-dose**, A protein catabolism 35%, B apoptosis 2%, C protein synthesis 9%, D RNA metabolism/transport 7%, E cell adhesion 2%, F DNA repair 2%, G nuclear transport 2%, H vesicle transport 2%, I unknown 26%, J development 2%, K transcription 3%, L protein folding 2%, M cell signaling 3%, N cytoskeleton 3%.

Figure 3.

Mass measurements benefit from software recalibration based on initial, confident identifications. The plot on the left shows precursor ion measurements from a Thermo LTQ Orbitrap (with identifications made by two different programs called ByOnic and PreView); the one on the right shows fragment ions from a QTOF instrument. In each case, the median error can be reduced by a factor of about 4 with software recalibration; the red line in the plot on the right shows a quadratic regression curve, although other fitting procedures for recalibration are available including nonparametric methods when large numbers of mass spectral data points are available [18].

Figure 4.

Conceptual workflow for computational analysis in the biomarker pipeline. The separation between dimension reduction and discriminant analysis is not as complete as shown in the figure, in that some methods incorporate aspects of both. Subsets of these approaches can be used.

Figure 5. Calculated dependence of statistical power on effect size for paired samples of sizes $n=15$ and 30, using significance levels of 0.05 and 0.01. The red circles are at 80% statistical power for a significance level of $p=0.01$.