Published in final edited form as:

Curr Protoc Bioinformatics. 2013 March ; 0 13: Unit–13.21. doi:10.1002/0471250953.bi1321s41.

Proteomics and the Analysis of Proteomic Data: 2013 Overview of Current Protein-Profiling Technologies

Can Bruce, Kathryn Stone, Erol Gulcicek, and Kenneth Williams

W.M. Keck Foundation Biotechnology Resource Laboratory, and Molecular Biochemistry and Biophysics Department, Yale University, New Haven, CT 06511

Abstract

Mass spectrometry has become a major tool in the study of proteomes. The analysis of proteolytic peptides and their fragment ions by this technique enables the identification and quantitation of the precursor proteins in a mixture. However, deducing chemical structures and then protein sequences from mass-to-charge ratios is a challenging computational task. Software tools incorporating powerful algorithms and statistical methods improved our ability to process the large quantities of proteomics data. Repositories of spectral data make both data analysis and experimental design more efficient. New approaches in quantitative and statistical proteomics make possible a greater coverage of the proteome, the identification of more post-translational modifications and a greater sensitivity in the quantitation of targeted proteins.

Keywords

LC-MS/MS; shotgun proteomics; search engines; protein inference engines; de novo sequencing; spectral counting; spectral library; label-free quantitation; labeled quantitation; data repository; Selective Reaction Monitoring; data-independent analysis; data exchange format; biomarker; network; pathway; post translational modification; phosphorylation

Introduction

Mass spectrometry (MS) has become one of the most powerful technologies in the analysis of the proteome. These techniques have displaced traditional techniques of biochemistry by enabling the identification and quantitation of proteins from complex mixtures. Because protein structure can be much more heterogeneous than DNA structure, experimental techniques of proteomics have evolved differently from DNA sequencing techniques. Whereas DNA sequencing techniques are general and scalable, mass spectrometric techniques have relied on sampling of partial information and making probabilistic inferences based on incomplete information.

Mass spectrometry is a sensitive method and measures the total mass (total mass (m) divided by charge (z), or m/z to be more exact) of a molecule (e.g. peptide or protein). Fragmentation of peptides and small proteins in the instrument enables protein identification based on measuring the masses of these fragmentation products and database searches. The quantitative challenges of mass spectrometry arise from the large amount of data produced by the fragmentation approach, as well as the difficulties associated with converting these data into protein sequences and protein quantity.

Current techniques of proteomics can be broadly divided into discovery and quantitative validation methods. While initial studies in a research project typically aim to identify the components of a protein sample, later studies often aim to quantitate the known proteins

(called protein profiling) either relative to a control sample, or on an absolute scale, for the purpose of establishing norms or identifying outliers.

DISCOVERY PROTEOMICS

The first step of protein profiling is the detection of peptides (and thereby, proteins) in a sample. Most projects start with a discovery phase where the investigator is simply interested in knowing what proteins are present while later phases may involve the quantitation of proteins on an absolute scale or in comparison to those in another sample. The discovery phase involves establishment of experimental protocols that result in the identification of the proteins in a particular sample (e.g., plasma). When the proteins of interest are likely to be present in low abundance, the probability of their detection may be increased by carrying out fractionation steps that either enrich for the protein class of interest (e.g., DNA affinity chromatography to enrich for transcription factors) or that deplete highly abundant proteins (e.g. immunoaffinity chromatography to deplete the 2–20 most abundant plasma proteins)..

Proteome analysis can be performed in two different strategies, called bottom-up and top-down approaches (Yates et al., 2009). In the bottom-up approach, a crude protein mixture is digested by a protease, followed by fractionation in one or two dimensions by liquid chromatography, and then MS analysis. Once the peptide sequences are determined, the protein sequences can be assembled from the peptide sequences. The assembly process can be complicated due to homologous proteins and alternatively splicing products. Also, labile PTMs, such as phosphate groups, may be lost, making their identification more difficult. This approach is commonly referred to as shotgun proteomics because of its similarity to shotgun sequencing as used in genomics (Claassen, 2012). In proteomics, the shotgun approach matches the observed parent ion with a predefined mass tolerance and its fragment ions with a predefined tolerance to a theoretical fragment ion containing database (usually derived from genomics) about the full protein sequences. In the top-down approach, intact proteins are characterized by MS without proteolysis. This type approach has the advantage of providing greater sequence coverage, resolution of sequence ambiguities and preservation of PTMs. However this strategy is technically more challenging, so is not widely used. Some of the shortcomings of the bottom-up approach have been overcome by the development of computational methods such as probabilistic scoring techniques, and use of data repositories to facilitate the identification of peptides (see below).

Discovery phase in proteomics is commonly referred to as shotgun proteomics because of its similarity to shotgun sequencing as used in genomics (Claassen, 2012); both approaches identify random fragments of a large sequence and assemble them by using prior knowledge (usually derived from genomics) about the full protein sequences.

Except for the most simple of samples, there is usually a need to simplify the sample into fractions which can be accomplished by a number of classical separation methods. Reversed phase liquid chromatography coupled to tandem mass spectrometry is most commonly used, although very complex samples or specialized needs may require the use of two successive steps of separation based on different physical principles, such as electrophoresis and isoelectric focusing; or peptide properties, such as cation exchange (i.e., charge) followed by reversed phase (i.e., hydrophobicity) chromatography (Link et al., 1999).

Mass spectrometric identification of peptides and hence, proteins is based on determining the mass of the peptides and then the mass of fragment ions derived from them (called tandem mass spectrometry or MS/MS). Typically, proteins are cleaved with sequence-specific proteases (such as trypsin) so optimal size peptide can be generated and their masses (actually, their m/z values) are determined. In tandem mass spectrometry, each

tryptic peptide ion whose m/z value is measured is then fragmented by collision with gas molecules to generate a series of product ions that are broken at various positions along the peptide backbone. Depending on which type of bond is broken along the peptide backbone, the resultant fragment ions are denoted as a_N , b_N , or c_N if they contain the amino-terminus of the precursor peptide, and x_N , y_N or z_N if they contain the carboxy-terminus, N being the number of residues in the fragment. The m/z value of the precursor ion and the set of masses corresponding to the resulting fragments is often sufficient to uniquely identify the sequence of the starting peptide, using database search algorithms. This is particularly true if the genome sequence of the organism from which the sample was obtained is available. Similarly, one or several peptides, whose identity is established by tandem mass spectrometry, are often sufficient to identify the parent protein. The peptide and protein identification search engines are described in greater detail below.

Sample Preparation—Biological samples are typically highly complex, so there usually is a need to reduce this complexity before starting mass spectrometry. Furthermore, because whole proteins are generally too large to be fragmented [e.g., by collisionally-induced dissociation (CID)] in a tandem mass spectrometer, they are fragmented by a protease (typically trypsin) before analysis. In some workflows, the proteins are fractionated by gel electrophoresis or chromatography, before being fragmented, while in others the fragmentation is done first, then the peptide mixture is fractionated, usually by reverse-phase liquid chromatography. When combined with the mass spectrometric analysis, this is called LC-MS/MS. The former approach is usually taken to identify variant forms of proteins as it is not possible to determine which peptides occur together in a protein once the protein is proteolyzed. However, the later approach is experimentally simpler and sufficient if the objective is just to identify or quantitate a particular protein or one of its modified forms.

The separation of peptides on an LC column not only serves to simplify the mixture to be analyzed, but also provides additional information that may be helpful, for instance, in distinguishing isobaric peptides that have the same mass. The fact that multiple ion fragments that are consistent with a certain parent ion are seen in adjacent fractions while fragments consistent with another parent ion do not show such a pattern would support the former peptide's identity. There are programs such as SSRCalc (Krokhin et al., 2006) and ELUDE (Moruz et al., 2012) that can predict with high specificity the retention time of a peptide on a reverse-phase column, and these can be used to resolve ambiguities. Additionally, in shotgun experiments, the differences between predicted and observed retention times can be used to detect incorrect peptide identifications [see references 17–21 in (Moruz et al., 2012)].

Search Engines—Data from a shotgun MS experiment is used to assign a sequence to each mass that is detected. This can be accomplished by either 1) matching the actual spectra to theoretical spectra based on the protein sequence database for the organism being studied (Database-dependent search engines), or 2) matching it to actual spectra saved in a repository (spectral matching engines) of the measured masses with sequence databases, or 3) de novo sequencing (which actually compare the observed mass data and their difference data with the theoretical mass of every possible sequence). PEAKS-DB uses a combination of all three approaches while sequence tagging approaches use database-dependent and de novo sequencing (Xu and Ma, 2006).

Database-dependent search engines (such as MASCOT (Perkins et al., 1999), SEQUEST (Eng et al., 1994a), PEAKS (Xu and Ma, 2006), X!TANDEM (Craig and Beavis, 2004), OMSSA (Geer et al., 2004), ProBID (Zhang et al., 2002), Phenyx (<http://www.genebio.com/products/phenyx/>), SONAR (Field et al., 2002) compare observed spectra with database-predicted spectra and assign probabilities that the observed peaks correspond to peaks that

were theoretically predicted based on known peptide sequences. The performance of these software have been compared (Nesvizhskii et al., 2007). They differ from one another in terms of speed, accuracy, sensitivity and false positives (i.e., incorrectly identified peptides). Because they work by differing algorithms, using multiple algorithms has been shown to yield fewer false positives, better coverage and higher confidence (Nesvizhskii et al., 2007).

The various search engines use different scoring algorithms and have varying advantages and disadvantages. All search engines start by filtering noise peaks, extracting m/z values and assigning charge values to the peaks. SEQUEST uses a cross-correlation function to match an observed with a theoretical spectrum. Mascot (based on an earlier search engine MOWSE (Pappin et al., 1993)) calculates the probability that an observed peak matches a theoretical one by chance, (Perkins et al., 1999). X!Tandem starts by creating a sub-list of proteins that have at least one minimal size tryptic peptide that matches the data. Because this sublist is significantly smaller than the full list, subsequent analysis of the data for incomplete cleavages or post-translational modifications can be done rapidly; a score is calculated for all matches to the database, then the p-value is calculated for the highest scoring match based on the hypergeometric distribution (Craig and Beavis, 2004). SEQUEST (Eng et al., 1994b) (Thermo Finnigan), MASCOT (Matrix Science), and PEAKS (Bioinformatics Solutions Inc.) are proprietary programs Open Mass Spectrometry Search Algorithm [OMSSA] (Geer et al., 2004).

X!Tandem, and Andromeda (Cox et al., 2011) are open-source software and have the advantage of being customizable or extendable. Indeed, parallelized versions of SEQUEST (Sadygov et al., 2002) and X!Tandem (Bjornson et al., 2008), run on a computing cluster and achieves speed gains by breaking up the computational task among many processors that run in parallel (Bjornson et al., 2008). Search engines calculate the probability that an observed peak matches a theoretical one by chance, using a database of invalid peptide sequences (a “decoy database”, generated by reversing or somehow randomizing the sequences of protein sequences for the organism being studied) (Elias and Gygi, 2010; Kall et al., 2008; Moore et al., 2002).

Spectral matching depends on the existence of a large number of tandem MS spectral data sets that may have been saved in the same laboratory, or are available in public repositories (see section on Repositories, below). As more and more data are accumulated for a particular organism’s proteome, this can become a viable approach to identifying peptides. The availability of spectral libraries is also a critical aspect in the development of targeted proteomic techniques such as SRM or SWATH (see section on Targeted Proteomics below). SpectraST (Lam et al., 2007), BiblioSpec (Frewen and MacCoss, 2007), and MSPepSearch (<http://peptide.nist.gov/>) are among software that can create and search spectral libraries.

De novo sequencing is the identification of a peptide that is carried out without the use of either a sequence database or spectral library. Database-independent approaches for de-novo sequencing use knowledge about the mechanism of ion fragmentation used in MS (such as collision induced dissociation (CID), electron-capture dissociation (ECD) or electron-transfer dissociation (ETD)). High mass-accuracy is critical in the application of this approach because some mass differences are nearly identical for certain amino acid combinations (eg., W and EG differ by 0.015 Da) Considering that there may be other peaks in the vicinity of a peak, a mass resolution of a few ppm is needed to resolve such ambiguous mass differences. This sensitivity is also adequate for searching complex databases containing modifications, polymorphisms, and their combinations (Mann and Kelleher, 2008). Software to identify the sequence directly from the MS/MS data has been developed (Hughes et al., 2010). PEAKS (UNIT 13.10; Bioinformatics Solutions, Inc.), PepNovo (Frank, 2009), and SHERENGA (Dancik et al., 1999) are some of the more

commonly used programs. One approach to accomplish this is the spectrum graph (Chen et al., 2001), where the spectral data is represented as nodes and edges of a network, each node corresponding to an observed m/z and each edge corresponding to the mass of an amino acid that accounts for the mass difference between two nodes. The various algorithms that use this approach differ in terms of how they deal with gaps in the sequence, or the likelihoods of different fragmentation events, but end by reporting the highest scoring path that connects the nodes of the network. The performance of these algorithms suffers when adjacent y- or b- ions are missing in the spectrum. The PEAKS algorithm does not use a network approach, instead it calculates a score for the top 10,000 peptide candidates corresponding to the spectrum. For each observed peak it calculates whether it corresponds to a y- or b-fragment based on the presence of neutral loss (water or ammonia) products nearby. For each candidate, the scores of the peaks that are consistent with it are summed and a confidence score is calculated. EigenMS models the links in the network as springs (Bern and Goldberg, 2006). A related software called RAId supplements this approach with a database search for homologous sequences (Alves et al., 2008). Combining the fragmentation data obtained from two different methods (CID and ECD/ETD) can significantly improve the confidence of the de novo sequencing (Savitski et al., 2005a; Savitski et al., 2005b; Zubarev et al., 2008).

Peptide sequence-tag dependent sequencing uses a combination of database-dependent and de novo sequencing. It starts by the identification of a two- or three-residue long peptide that can be identified with high confidence. The sequence tag is used to narrow down the search in the database. The sequence tag is extended by searching for additional fragments that match the last identified peptide plus one more amino acid (Mann and Wilm, 1994). The programs Gutentag (Tabb et al., 2003), InsPecT (Tanner et al., 2005) and Myrimatch (Tabb et al., 2007) are automated tagging programs. A peptide sequence tag may not be detectable if the spectrum is noisy.

The peptide sequencing approaches mentioned can be combined, with PEAKS-DB utilizing all three approaches described above.

Protein Inference Engines—Once peptides are identified, their sequence is used to assemble the sequences of the proteins in the original sample. This is not a straightforward exercise, because some peptide sequences may be common to multiple proteins due to alternative splicing or membership in homologous families. Furthermore, small proteins may not have as many identified peptides as large ones. Also, peptides differ in the confidence levels of their identification and that determines the confidence in the identification of the protein. Software called protein inference engines convert peptide level identification to protein assignments. Some common examples are ProteinProphet (Nesvizhskii et al., 2003), Empirical Bayes Protein (EBP) identifier (Price et al., 2007) and Mayu (Reiter et al., 2009).

Identification of peptides can suffer from a number of uncertainties that may, for instance, arise from instrument measurement uncertainties, unexpected chemical or genetic modifications in the sequence, and physicochemical phenomena that make certain fragments nearly undetectable. It is necessary to assign a probability value to the possibility that the observed data matches the identified peptide sequence by chance, and furthermore to correct this value for statistical issues arising from multiple comparisons. To estimate the false positive rates, a commonly used approach is to create a theoretical spectral library *in silico* from proteins whose sequences are not found in the proteome of the organism being studied. Matches of the observed spectra to sequences in this “decoy database” are used to calculate the false positive rate (usually 1%). Different search engines have been evaluated with the conclusion that they give complementary results and that results can be improved by

combining results from multiple search engines (Kapp et al., 2005); the topic has been reviewed more recently by (Nesvizhskii, 2010). While the statistics of peptide matches is relatively well understood, that of protein inference from peptide data is a more complex problem that is still being studied (see review (Claassen, 2012)). The errors at the level of peptide identification need to be propagated to the level of protein identification but this can be complicated by protein identifications supported by correct as well as incorrect peptide spectrum matches. Especially when very large protein data sets (including multiple members of protein families or simple variants caused by SNPs) are used, false discovery rates can be much higher than may be expected from naive assumptions. The MAYU software (Reiter et al., 2009) uses a hypergeometric model to estimate false positive rates in protein inference and due to its successful validation, has been incorporated into the Trans Proteomic Pipeline (see below) and been applied to the 2010 compilation of the human plasma proteome in PeptideAtlas (Farrah et al., 2011). The CEDAR scheme groups proteins identified from peptides into five groups according to the redundancy of their peptide sets. This allows each protein to be classified according to the ambiguity of its supporting peptide identifications (Farrah et al., 2011).

Experimental considerations in proteome cataloging—There are two strategies for maximizing the coverage of a proteome using an LC-MS/MS approach. The first is based on the assumption that there are too many proteins in the sample for them to be all identified in a single run. That is, because of the highly complex mixture of proteins that are present at detectable concentrations in the sample, during each run the instrument will only be able to identify a subset of all of these proteins. Therefore, repeating the experiment multiple times can be expected to sample an ever increasing proportion of the proteins that are in the sample at sufficiently high concentrations to permit identification. The number of repetitions needed can be estimated by modeling the progression of proteome coverage as an infinite Markov model (Claassen et al., 2009). A variation of this approach is to carry out the repeat analyses with an exclusion list of already identified peptides which helps to enable the mass spectrometer to focus on obtaining MS/MS data from peptides that have not yet been identified. An alternative and complementary approach is to analyze qualitatively different samples. Compiling a comprehensive list of proteins present in a complex proteome such as that obtained from a particular tissue or biological fluid (e.g., human plasma) requires the use of different sample preparation, fractionation, and analysis techniques; for the proteome of an organism (e.g., human proteome), different tissues and fluids also would need to be sampled. In both instances, it may be advantageous to repeat the studies using peptides that were generated by the use of different proteases. Because each technique yields different sets of proteins, a comprehensive proteome can be compiled only by combining data from many diverse, high-quality experiments (Farrah et al., 2011). Computational simulations have suggested improved experimental protocols in shotgun proteomics. Fractionation of proteins first before proteolysis and analysis of the peptides produces better coverage of the proteome and higher relative dynamic range, as compared to performing proteolysis first and then fractionating the peptides and analyzing the peptide fractions (Eriksson and Fenyo, 2007). Another advantageous strategy based on computational techniques is to use a directed proteomics approach (see section on SRM below): the identification of a peptide from high-quality and redundant peaks would suggest the presence of a protein and then a second run would be performed where peptides for the suggested protein are targeted for confirmation or detailed characterization (Schmidt et al., 2008; Zerck et al., 2009).

Computational approaches can be used to improve proteome coverage by identifying proteins with physico-chemical characteristics that are under- or over-represented. For example, when Brunner et al. determined that their provisional catalog of the *Drosophila* proteome was biased against small proteins, they fractionated their cell extracts by gel filtration to focus on the small-size proteins (Brunner et al., 2007).

When reporting the list of proteins, it is best to fix the protein false discovery rate to a reasonably low level like 1%, as there is a trade-off between the number of proteins identified and their confidence levels (Claassen, 2012). The use of sophisticated protein inference machines that focus on error minimization have been shown to work better than simple rules like avoiding protein identifications that are based on a single peptide identification (Claassen et al., 2012; Gupta and Pevzner, 2009; Reiter et al., 2009).

Data repositories—With the development of proteomics, a need has emerged for the creation of data repositories for data generated from proteomics experiments, for data exchange between collaborators, as well as from publication requirements of scientific journals. Although a number of repositories have been created over the years, several have stopped being updated. Such repositories vary in size and functionality depending on their user base and purpose.

Some large academic proteomics centers have developed their databases for the needs of their users. The Yale Protein Expression Database (YPED) is a Web-based laboratory information management system that allows the storage and display of proteomics data from various technologies, along with experimental metadata (Shifman et al., 2007). Other repositories are used for the archival or proteomics data. Among major data repositories are. Human Proteinpedia, Yeast Resource Center Public Data Repository, Yale Protein Expression Database (already mentioned above), PRIDE (UNIT 13.8), Global Proteome Machine, PeptideAtlas. Collections of tandem mass spectral libraries are available at GPM (Beavis, 2006), SpectraST (Lam et al., 2007), and BiblioSpec (Frewen and MacCoss, 2007).

Detection of post-translational modification—Protein sequences that are modified by post-translational modification (PTM) in vivo or artifactually in vitro (through oxidation etc.) present a challenge for the determination of peptide sequences. Especially if there are multiple candidate modification sites near each other, the localization of the actual one may be difficult. A number of programs are available to calculate the probability that a particular modification exists and to localize it. Phosphorylation is the most common PTM studied although other modifications such as ubiquitination, acetylation, nitrosylation and glycosylation are also important. A peptide containing a phosphorus atom can usually be identified simply from its fractional mass (Bruce et al., 2006) but its precise localization requires detailed analysis of its ion-fragment masses. AScore (Beausoleil et al., 2006) and Phosphoscore (Ruttenberg et al., 2008) are phosphorylation assignment programs that use SEQUEST output files, MascotDelta Score (Savitski et al., 2011) is an extension of the Mascot search engine, for phosphorylation site analysis. The InsPecT program can predict phosphopeptide sequences more accurately than SEQUEST or MASCOT by making use of a large training dataset (Tanner et al., 2005). PhosphoRS (Taus et al., 2011), Colander (Lu et al., 2008), APIVASE (Jiang et al., 2008) and DeBunker (Lu et al., 2007) are other programs that use various algorithms to validate phosphopeptides and localize phosphorylation sites. These programs can be used in conjunction with programs that can predict phosphorylation sites based on protein sequence. GPS (Zhou et al., 2004), ScanProsite (Gattiker et al., 2002), PhosphoELM (Dinkel et al., 2011), Phosphomotif finder (Amanchy et al., 2007), Predikin (Ellis and Kobe, 2011), NetPhosK (Blom et al., 2004), PredPhospho (Kim et al., 2004), KinasePhos (Wong et al., 2007), NetPhorest (Miller et al., 2008) and PostMod (Jung et al., 2010) are some programs that perform phosphor-site prediction based on already-known sites or kinase specificities. The challenges of phosphorylation analysis are applicable to some of the other PTMs.

QUANTITATIVE PROTEOMICS

A major task in most protein profiling projects is to quantitate the amount of each protein in a sample either relative to another sample, or in absolute terms (e.g., copy number/cell). The available techniques can be broadly divided into data-dependent and data-independent approaches. In data-dependent approaches, the selection of peptide fragments to be fragmented for MS/MS analysis is determined by pre-established criteria. In data-independent approaches, all peptides (within a pre-determined mass range) are fragmented indiscriminately. Data-independent approaches are relatively new but hold great promise for the future of quantitative proteomics. Advantages of this approach are that it potentially can identify all species that are present in a given sample at detectable concentrations in a single analysis and that the data can be retrospectively interrogated as improved software and in silico databases become available. The large majority of quantitative proteomics data have so far been generated by data-dependent techniques, which can be grouped into labeled and label-free approaches. Labeled quantitation relies on the principle of labeling multiple samples uniquely prior to processing them together up to the measurement step, when peptides or fragment ions originating from different samples can be distinguished. This approach has the advantage of minimizing error rates because, after the labeling step, experimental sources of variability apply equally to both samples. iTRAQ, ICAT, DIGE, SILAC/SILAM are labeled quantitation techniques. Label-free quantitation (LFQ) is performed on samples analyzed serially and suffers from uncertainties in estimating the concentration of each peptide, but benefits in its ability to run countless sample comparisons. It also has greater dynamic range and greater coverage of the proteome. With the improvement of computational methods in estimating concentrations, label-free techniques are becoming more popular.

Labeled quantitation—Most labeled quantitation approaches rely, as the name suggests, on the incorporation of a label into one or more of the samples to be analyzed. This can be accomplished at the metabolic-level in vivo, or by chemical or enzymatic derivatization in vitro. Derivatizing proteins with a fluorescent label also falls under this category, although it is used to identify differentially expressed proteins separated by 2-dimensional gel electrophoresis (DIGE).

In stable isotope labeling by amino acids in cell culture (SILAC) or mouse (SILAM), a metabolically labeled amino acid is incorporated into the growth medium of a cell culture, or fed to an animal, in order to metabolically label its proteins. Later, this sample is combined with a control sample containing a heavy stable isotope version of the same amino acid. After processing of the samples (purification or fractionation of proteins, proteolysis, and MS analysis) in parallel, those proteins that are differentially expressed can be identified. SILAC can be performed using different heavy stable isotopic forms of arginine to perform up to 5-plex analyses (Molina 2010).

Chemical derivatization of amino acid side chains is another labeling method. In the isotope-coded affinity tag (ICAT) approach, cysteines are derivatized with a reagent containing either zero or nine ^{13}C atoms, and also a biotin group that is used to purify the derivatized peptides (from AB Sciex Pte. Inc.). In subsequent MS analysis, modified peptides originating from either the control or treated sample can be quantified. Variations on this approach using other labeling reagents exist and have been reviewed (Bantscheff et al., 2007). Peptides can also be labeled on the N-terminal and the epsilon-amino group of lysines, with stable-isotope tagged reagents. Among the variants of this approach (ICPL, iTRAQ, TMT and others), the iTRAQ reagents have the advantage of having precisely the same mass until the peptide fragmentation step, after which samples originating from

different samples end up having different mass labels. Up to eight mass-labels can be processed simultaneously with this reagent.

Absolute quantitation—Spiking into the sample an internal standard peptide that is stable isotope-labeled can help quantify the levels of the protein containing the sequence of this peptide. In Selective Reaction Monitoring (SRM: see section below), a number of internal standards are introduced into the sample and the mass spectrometer is programmed to measure specifically the signal intensity of their peptide masses and one or more of their fragment ions for the labeled internal standard and also for the native, unlabeled form.

Label-free quantitation—Label-free quantitation strategies are based on one of two different principles: a) The quantity of a protein is proportional to the mass spectrometric signal of one or more of its peptides, and b) the number of fragment spectra identifying peptides of a given protein is proportional to the protein's quantity (spectral counting) (Liu et al., 2004). Comparing either of these metrics across samples yields a measurement relative to a control sample. Label-free quantitation is typically done in an LC-MS/MS system, and the combination of chromatographic retention time, peptide mass and fragment ion masses are used to uniquely identify peptides. Compared to labeled quantitation, label-free quantitation has the advantages that it has greater throughput of data, a higher sensitivity with less than 1 μ g total protein required (versus a typical $>1\mu$ g amount using labeled techniques) and is less expensive in terms of reagent costs. But, since samples are run in triplicate and blanks are required between runs to ensure no peptide carryover, a large amount of instrument time is needed.

A number of issues have necessitated the development of software for label-free quantitation. Because LC separation of successive samples occurs serially, depending on the instrument, there may be a need to align chromatographic peaks accurately, depending on the instrument's reproducibility. Software such as msInspect (Fitzgibbon et al., 2008) and polyAlign (Vahamaa et al., 2011) accomplish this. Another issue is the proportion of time to be spent in the survey (MS) mode, versus in the spectra acquisition (MS/MS) mode. A nano-LC column typically produces precursor ion width of 10–30 seconds and this limits how much time can be spent in total between these two modes. While quantitation needs to start with an identification step, after that there is trade-off between maximizing the number of peptides identified and their quantitation. During each cycle, the more time each peptide is measured, its quantitation is more accurate, but less time is left to identify new peptides.

Quantitation by peptide or spectral counting is based on the fact that because fragment ions have different signal intensities, their detectabilities vary widely and therefore the more precursor peptide there is, the more fragment ions will be detectable. However, the relationship between quantity and number of detectable spectra is protein-dependent. Spectral counting methods can be relatively sensitive, for example after normalizing for the number of peptides obtainable from a protein, 22 spectra was sufficient to detect a 1.4-fold quantity change in one protein (Zybailov et al., 2006).

When using the sum of the signal peak heights or areas under the curve for the (isotopic envelope) peaks corresponding to a peptide, there is good correlation between this value and the quantity of the source protein. Still, because peptides vary in their signal strength, accurate quantitation requires the use of the relative intensities of several peptides. A number of factors contribute to the error in assessing protein quantity from peptide signal: a) background noise, b) interfering signals from co-eluting peptides of similar m/z , c) detector saturation. In general, strong signals are more accurate, so using an intensity-weighted average, or performing a linear regression on all pairs of signal values for peptides of a given protein and obtain a slope value, increases accuracy. Another challenge with label-free

quantitation is that some peptides can be shared by multiple proteins. Several computational approaches can be used to deal with this challenge, but the best approach distributes the contribution of each shared peptide among multiple proteins based on the number of unique peptides per protein (Zhang et al., 2010).

This requirement limits the proteins that can be accurately quantitated to those that have a sufficient number of peptides observed. The development of targeted assays like SRM should help to overcome this limitation, because once a protein has a reasonable number of spectra observed, the optimal ones can be selected for quantitation.

A number of commercial and open-source software are available for label-free proteomics data analysis. Among the commercial ones are: Nonlinear Dynamics Progenesis LC-MS (<http://www.nonlinear.com/products/progenesis/lc-ms/overview/>); Decyder MS (GE Healthcare; <http://www5.gelifesciences.com/>); SIEVE (Thermo Electron <http://www.thermo.com/>); Elucidator (Rosetta; <http://www.rosettatabio.com/>); ProteinLynx (Waters; <http://www.waters.com/>). Each of these quantify peak intensities, with Elucidator and Progenesis LC-MS also providing spectral count analysis. Among the open-source software are MapQuant (Leptos et al., 2006), MZmine (Pluskal et al., 2010), MsInspect (Fitzgibbon et al., 2008), OpenMs (Sturm et al., 2008) (this is a software framework to develop components of TOPPS), MSight (Palagi et al., 2005), SuperHirn (Mueller et al., 2007), and PEPPeR (Jaffe et al., 2006). Spectral counting software is reviewed in detail in (Lundgren et al., 2010).

Selective Reaction Monitoring—Selective Reaction Monitoring (SRM; (Lange et al., 2008)) also known as Multiple Reaction Monitoring (MRM) (when the instrument is set to monitor more than a single mass per acquisitions) is a technique used to measure a limited number of proteins in a complex background in multiple samples. By selecting a small number (usually 2–5) of precursor ions and their fragments, the presence and quantity of a target protein can be accurately determined. This technique allows for the accurate measurement of relative amounts of a given peptide among multiple samples, or, if a heavy isotope-labeled version of the monitored peptide is spiked in an internal control, the absolute quantitation of the peptide. The most difficult part of this assay is identifying a combination of peptides and their corresponding fragment ions that are 1) unique to each protein of interest, 2) whose signal does not overlap with those from background proteins, and 3) give consistent and strong signal.

The selection of peptides and their fragment ions (“transitions”) is the most challenging aspect of this technique and it relies heavily on computational resources. Selection of the appropriate transitions is greatly facilitated by prior observation of peptides belonging to the protein of interest. Laboratories that have carefully compiled large amounts of spectral data can positively leverage their prior data to efficiently select peptides that satisfy the above requirements. There are also public repositories where proteomic data can be deposited to assist in the development of SRM assays. For example Peptide Atlas (Deutsch, 2010), MRMatlas (Picotti et al., 2008), EBI’s PRIDE (Vizcaino et al., 2010). An alternative approach to developing SRM assays is to use theoretical predictions of MS signal strength, which are based on various physical characteristics of peptides, to select the optimal transitions to monitor (Mallick et al., 2007). The SRMatlas (Picotti et al., 2008) contains spectrum libraries for a large number of crude peptides that were synthesized on a relatively massive scale (Sternagachis et al., 2011). Such SRM assay atlases can be expected to make the measurement of protein levels a more routine undertaking.

SRM studies present computational challenges both in the design of the method and also in the analysis of the data. One tool called ATAQS consists of software modules that carry out

all the steps in designing, managing, analyzing, and validating an SRM assay. The modular design of this tool allows investigators to go through some or all of the process using this tool. This open source system works with a java-based program that can be run on a Firefox Web-browser. The prediction of the ionization properties of tryptic peptides is an area of ongoing research. For lower abundance proteins, the MIDAS approach assists in identifying optimal SRM targets. Optimal transitions and transition validation can be guided by interfaces (for example, MRmaid (Mead et al., 2009), SRMatlas interface (www.srmatlas.org), MRMer (Martin et al., 2008) and MaRiMba (Sherwood et al., 2009) to identify those peptides with the highest signal and the lowest probability of having an interfering signal. Another tool is SRMCollider, which calculates unique ion signatures for a given proteome (Rost et al., 2012).

An intermediate step between discovery and SRM validation is to target predicted peptides to verify their presence in a verification set of samples (Jaffe et al., 2008). Retention times and characteristic, relative intensities of the transitions of each peptide are determined at this verification step and are used to design a more reliable SRM assay. An analogous method is MIDAS (multiple reaction monitoring-initiated detection and sequencing).

SRM analysis can be applied to the detection of phosphorylation and other PTMs. For example, the MIDAS approach allows the detection of hypothesis driven phosphorylation at specific sites. For each peptide containing a possible site, the mass spectrometer is set to detect that peptide's mass in its phosphorylated form; the diagnostic ion for the identification of a phosphoserine or phosphotyrosine is the loss of a H_2PO_4 group (98 Da); for tyrosine phosphorylation, the diagnostic ion is the generation of an immonium group (216 Da).

A standardized data format called TraML, has been devised to keep track of the steps in the development of an SRM assay. Once the assay is developed, the file format can also be used for the exchange of this information between the various software that are used at various steps in the development and application of a particular assay (Deutsch et al., 2012). It can also be used to store the information in appropriate repositories.

Data-independent approaches—The selection of a peptide for further fragmentation is often dictated by data derived during the analysis. For example, ions that have already been selected for fragmentation are often excluded dynamically, to maximize the number of peptides identified. The number of peptides that can be analyzed by tandem mass spectrometry is often limited during an LC elution, and instruments are often set to select the most abundant co-eluting peptides for fragmentation. In recent years an alternative approach has been developed in which fragmentation is performed independently of the peptide MS data that is obtained “on-the-fly”. In this approach product ions are not associated with specific precursor ions.

While indiscriminately fragmenting all peptide ions generates intractable complexity from samples containing thousands of proteins, fragmentation of peptides within a pre-determined m/z and LC retention time range results in a manageable number of fragment ion spectra. In data independent acquisition (DIA) approaches (Dong et al., 2007; Venable et al., 2004), PAcIFIC (Panchaud et al., 2009) fragments resulting from precursor ions within a window of 10–25 Da m/z range are measured successively, over a desired m/z range (like 400–1400). As the scan of each window typically takes in the order of 100 ms, the LC elution can be analyzed by this method every 20–30 sec, which is sufficient to resolve most typically occurring peaks. The resulting data can be queried against pre-existing spectral libraries for identification of their precursors. DIA data can be analyzed by the use of software such as mProphet (Reiter et al., 2011) or Skyline (MacLean et al., 2010). Currently, software tools

to exploit the full potential of DIA are still being developed. It is to be expected that the wealth of information in DIA data can be exploited using the knowledge of an organism's proteome. Spectral libraries for a number of organisms such as yeast, human, etc. (Bodenmiller et al., 2008) are available (<http://www.peptideatlas.org/speclib/>), and as more such spectral libraries become available, this technique can be expected to be applicable to a wider selection of proteomes.

Statistical Issues—Proteomics experiments generate large amounts of data that need to be processed, first to identify proteins with confidence, then to quantify them within a reasonable error rate. In experiments that compare two conditions, one often needs to know whether the observed change in expression of a particular protein is statistically significant. More generally, there is often a need to compile a list of proteins that show differential expression as a result of the treatment or other variable being studied. The proper statistical treatment of these tasks is critical in reaching correct biological insights.

Quantitative proteomics data needs many steps of processing before statistically valid conclusions can be reached. Raw data is often log-transformed because this results in a normal distribution. Replicates need to be normalized to each other and then to the control sample replicates. Systematic errors need to be corrected for either computationally (e.g., mixing errors in multiplexed experiments) or experimentally (e.g. by label swapping in labeled quantitation). Data that may be missing in certain samples that are to be compared may need to be imputed based on the values of replicates or by comparison to the relative level of expression of other related proteins. Data outliers need to be identified by various standard statistical methods. Once identified, then a decision must be made with respect to ascribing these outliers either to an experimental “glitch” or to a potentially important biological phenomenon, like post-translation modification of a particular residue [e.g., thus accounting (for instance) for the relatively low intensity of the corresponding non-modified peptide transition in an SRM assay].

Detection of differential protein expression depends on an assessment of the probability that an observed change in expression results from “random” factors related to measurement methods, rather than from a true biological effect. Replicate samples are required to establish the variability and precision of the measured quantities. Statistical tests (t-test, local-pooled error test and non-parametric tests) with varying degrees of statistical power have been proposed to assess the significance of the observed change (Jain et al. 2003; Jain et al. 2005, Zhang et al.2006). Because many proteins are quantitated, multiple testing errors need to be considered as well. The family-wise error rate (FWER) is the rate that one true negative is erroneously called positive; the false discovery rate (FDR) is the rate that true negatives are erroneously called positive within a list. FDR rates are less conservative as compared to FWER rates which thus increase statistical power at the cost of also increasing the rate of false positives. An adjusted p-value can be calculated to reject a hypothesis when multiple comparisons are conducted. This topic has been reviewed by (Dudoit et al., 2003).

In the case of label-free quantitation using spectral counts, each protein's spectral count divided by the sum of the spectral counts of all proteins detected in the sample can be used to assess differential expression. Three tests, Fischer's exact test, AC test and the G-test, were found to work best in testing the significance of differences in expression of individual proteins between two samples, and the G-test was found to be the computationally simple and applicable to multi-sample comparisons (Zhang et al., 2006).

Biological perspectives—Once a list of proteins are identified from a proteomics experiment (differentially expressed between two conditions, present in a certain biological compartment, or showing the same temporal expression pattern, for example) a number of

analyses can be done to provide biological perspective. Bioinformatics databases contain annotations on proteins, such as the chromosomal location of their genes, publications relating to them, diseases in which their expressions was found to be altered, other proteins they interact with, etc. The membership into any one annotation group can be evaluated by asking whether members of the list are seen in that group more frequently than can be expected by chance. Fischer's Exact Test (equivalent to the hypergeometric test) is used to assign a p-value to the enrichment, although some investigators prefer a modification of this test because it can declare a single observation to be statistically significant (Ease score: (Hosack et al., 2003)). The DAVID web site can carry this analysis one step further, by clustering proteins into enrichment groups, such that a group of proteins can be said to be enriched in two or more annotation terms belonging to different databases (such as a cellular component and a molecular function ontology term) (Huang et al., 2007).

Many proteins are known to interact with each other: an enzyme-substrate relationship (e.g., a protein phosphorylation event), gene regulation step (a transcription factor regulating the expression of the mRNA coding for a protein), or a binding event are some examples of interactions that can be used to place proteins in a network. It is possible to construct networks of interacting proteins in which members of a list of differentially expressed protein are statistically enriched. Other proteins that are in this network can be hypothesized to be causally required to account for the changes seen in the whole network. A number of network-generating programs that interconnect a given list of proteins are available (Ingenuity Pathway Analysis (Ingenuity), Metacore (GeneGo), Pathway Studio (Ariadne Genomics)), they differ from each other in the algorithms they use and in the content of their interactions databases. Commercial companies claim to have manually annotated databases that are richer than public-domain databases.

Biomarker discovery—Proteins identified in comparisons of patient and healthy controls samples often yield a list of differentially expressed proteins. Because of the high biological variability (e.g., between two different control patients) in the proteome of body fluids typically used in biomarker studies (plasma, urine, etc.) and the low number of patients typically assayed in the discovery phase of biomarker studies, biomarker candidates need to be confirmed in a second, larger group of samples. It is also essential to either confirm that the candidate proteins are expressed stably over time so that their levels are consistently different between the control and diseased population or to determine the “time window” since, for example, disease onset that a particular biomarker is differentially expressed. Furthermore, an ideal biomarker needs to distinguish not only a diseased person from healthy ones, but also from persons who have other types of diseases. The SRM technique is especially appropriate for the validation step because an inclusion list can be created from the list of peptides that uniquely identify candidate proteins to economically measure their levels in the large numbers of samples that need to be assayed to validate putative biomarkers. It has been posited that MRM protein assays “may soon surpass ELISAs for the rapid development of clinically useful, multiplexed protein assays” (Domanski et al., 2012). The use of mass spectrometry in biomarker discovery has been reviewed in detail (Hawkrige and Muddiman, 2009).

Integrated software for proteomics assays—The various proteomics approaches outlined above require a series of data processing steps. Software that streamlines these analyses have been developed by companies and academic labs.

Label-free analysis requires a number of data processing steps. 1) multiple LC/MS data sets need to be normalized for sample load, 2) peaks need to be aligned among multiple LC runs, 3) peak detection, 4) peak integration, 5) peak quantitation, 6) peak matching, 7) statistical analysis. A review article (Zhu et al., 2010) summarizes the open-source and commercial

software available to process label-free proteomics data. Some prominent data analysis systems are MaxQuant (Cox and Mann, 2008) and Trans Proteomic Pipeline (Keller et al., 2005) (Pedrioli, 2010)

Data exchange formats—Because there are many good software that are good at accomplishing specific tasks in a data processing pipeline, it has become important to many researchers that these software can be integrated with each other in such a way that the output of one can readily serve as the input of another one. Proprietary or difficult-to-parse outputs were a challenge to achieving this goal but over the years, a number of standard, XML-based file formats have been devised within the research community. For example mzXML format developed by the Sashimi project [<http://sashimi.sourceforge.net/>] and the mzData developed by the Human Proteome Organization (<http://hupo.org>) allow peptide identifications by different software to be combined by integrative programs such as by Scaffold (Searle, 2010) or PEAKS DB (Zhang et al., 2012). At this point, application-specific file formats such as Mascot dat and mgf files, SEQUEST result and datfiles, MS2 and DTASelect formats can be converted into XML-based formats such as mzXML (Pedrioli et al., 2004), mzML (Turewicz and Deutsch, 2011), pepXML (Keller et al., 2005), protXML (Keller et al., 2005), mzIdentML (Eisenacher, 2011), mzQuantML(<http://www.psidev.info/mzquantml>), TraML(Deutsch et al., 2012), X! Tandem XML(Craig and Beavis, 2004), PRIDE XML(www.ebi.ac.uk/pride/).

A Minimum Information About a Proteomics Experiment (MIAPE(Taylor et al., 2007)) specification called MIAPE-Quant is being developed at the time of this writing by the HUPO Proteomics Standard Initiative(www.psidev.info/). Its aim is to ensure that sufficient information about an experiment is stored “to allow reproducible repetition of such an experiment and/or to allow judgment of the experimental and analysis quality regarding the experimental design and samples used, the analysis parameters and the results.”

Conclusion

While protein identification from classical biochemical techniques (such as antibody or enzymatic activity-based) is essentially qualitative, mass spectrometry-based protein identification is highly computational because of the stochastic and noisy aspects of the technology. The great strides that made in the field have resulted not only from better instrumentation and more powerful computers but also from better algorithms and statistical tools, as well as our ability to take advantage of the accumulated data(as those from spectral libraries). At the lowest level, matching observed peaks to theoretical ones is associated with uncertainty that requires techniques of signal processing. But those molecular species that are identified with high confidence represent an incomplete sampling of all that are potentially present. The identification of proteins depends on the propagation of the uncertainties in the identification of ion fragments and their precursor ions in ways that are still being studied. Furthermore, the quantitation of the levels of any one protein is affected by many sources of random errors whose statistical treatment is not completely clear. The final picture that emerges about changes in a proteome as a result of a treatment is a probabilistic one where the amount of detail depends on how much uncertainty we are prepared to tolerate. There are many challenges left: the fraction of the fraction of the open reading frames in genomes that have been detected by mass spectrometry is still small, the cataloging and dynamics of post translational modifications has barely started, and very little has been studied of the quantitative changes that result in the course of biological processes. Further improvements in computational methods of mass spectrometry can be expected to contribute to advances in the analysis of proteomic data.

Acknowledgments

This work was supported in part with Federal funds from NIH grants DA018343, DK090744 and UL1 RR024139.

Literature Cited

- Alves G, Ogurtsov AY, Kwok S, Wu WW, Wang G, Shen RF, Yu YK. Detection of co-eluted peptides using database search methods. *Biology direct*. 2008; 3:27. [PubMed: 18597684]
- Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. *Nature Biotechnology*. 2007; 25:285–286.
- Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*. 2007; 389:1017–1031. [PubMed: 17668192]
- Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*. 2006; 24:1285–1292.
- Beavis RC. Using the global proteome machine for protein identification. *Methods Mol Biol*. 2006; 328:217–228. [PubMed: 16785652]
- Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *Journal of computational biology: a journal of computational molecular cell biology*. 2006; 13:364–378. [PubMed: 16597246]
- Bjornson RD, Carriero NJ, Colangelo C, Shifman M, Cheung KH, Miller PL, Williams K. X!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *Journal of Proteome Research*. 2008; 7:293–299. [PubMed: 17902638]
- Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*. 2004; 4:1633–1649. [PubMed: 15174133]
- Bodenmiller B, Campbell D, Gerrits B, Lam H, Jovanovic M, Picotti P, Schlapbach R, Aebersold R. PhosphoPep—a database of protein phosphorylation sites in model organisms. *Nature Biotechnology*. 2008; 26:1339–1340.
- Bruce C, Shifman MA, Miller P, Gulcicek EE. Probabilistic enrichment of phosphopeptides by their mass defect. *Analytical Chemistry*. 2006; 78:4374–4382. [PubMed: 16808444]
- Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O, Lee H, Pedrioli PG, Malmstrom J, Koehler K, Schrimpf S, Krijgsvelde J, Kregenow F, Heck AJ, Hafen E, Schlapbach R, Aebersold R. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nature Biotechnology*. 2007; 25:576–583.
- Chen T, Kao MY, Tepel M, Rush J, Church GM. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology: a journal of computational molecular cell biology*. 2001; 8:325–337. [PubMed: 11535179]
- Claassen M. *Molecular & Cellular Proteomics: MCP*. 2012. Inference and validation of protein identifications.
- Claassen M, Aebersold R, Buhmann JM. Proteome coverage prediction with infinite Markov models. *Bioinformatics*. 2009; 25:1154–1160. [PubMed: 19477982]
- Claassen M, Reiter L, Hengartner MO, Buhmann JM, Aebersold R. Generic comparison of protein inference engines. *Molecular & Cellular Proteomics: MCP*. 2012; 11:O110 007088.
- Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*. 2008; 26:1367–1372.
- Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*. 2011; 10:1794–1805. [PubMed: 21254760]
- Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004; 20:1466–1467. [PubMed: 14976030]

- Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology: a journal of computational molecular cell biology*. 1999; 6:327–342. [PubMed: 10582570]
- Deutsch EW. The PeptideAtlas Project. *Methods Mol Biol*. 2010; 604:285–296. [PubMed: 20013378]
- Deutsch EW, Chambers M, Neumann S, Levander F, Binz PA, Shofstahl J, Campbell DS, Mendoza L, Ovelleiro D, Helsen K, Martens L, Aebersold R, Moritz RL, Brusniak MY. TraML--a standard format for exchange of selected reaction monitoring transition lists. *Molecular & Cellular Proteomics: MCP*. 2012; 11:R111 015040.
- Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F. Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Research*. 2011; 39:D261–267. [PubMed: 21062810]
- Domanski D, Percy AJ, Yang J, Chambers AG, Hill JS, Freue GV, Borchers CH. MRM-based multiplexed quantitation of 67 putative cardiovascular disease biomarkers in human plasma. *Proteomics*. 2012; 12:1222–1243. [PubMed: 22577024]
- Dong MQ, Venable JD, Au N, Xu T, Park SK, Cociorva D, Johnson JR, Dillin A, Yates JR 3rd. Quantitative mass spectrometry identifies insulin signaling targets in *C. elegans*. *Science*. 2007; 317:660–663. [PubMed: 17673661]
- Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*. 2003; 18:71–103.
- Eisenacher M. mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. *Methods Mol Biol*. 2011; 696:161–177. [PubMed: 21063947]
- Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol*. 2010; 604:55–71. [PubMed: 20013364]
- Ellis JJ, Kobe B. Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. *PloS One*. 2011; 6:e21169. [PubMed: 21829434]
- Eng J, McCormack AL, Yates JR. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry*. 1994a; 5:976–989.
- Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994b; 5:976–989.
- Eriksson J, Fenyo D. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nature Biotechnology*. 2007; 25:651–655.
- Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmstrom J, Ossola R, Watts JD, Lin B, Zhang H, Moritz RL, Aebersold R. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Molecular & cellular proteomics: MCP*. 2011; 10:M110 006353.
- Field HL, Fenyo D, Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*. 2002; 2:36–47. [PubMed: 11788990]
- Fitzgibbon M, Law W, May D, Detter A, McIntosh M. Open-source platform for the analysis of liquid chromatography-mass spectrometry (LC-MS) data. *Methods Mol Biol*. 2008; 428:369–382. [PubMed: 18287783]
- Frank AM. Predicting intensity ranks of peptide fragment ions. *Journal of proteome research*. 2009; 8:2226–2240. [PubMed: 19256476]
- Frewen, B.; MacCoss, MJ. Using BiblioSpec for creating and searching tandem MS peptide libraries. In: Baxevanis, Andreas D., et al., editors. *Current protocols in bioinformatics*. Vol. Chapter 13. 2007. p. 17
- Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*. 2002; 1:107–108. [PubMed: 15130850]
- Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. *Journal of Proteome Research*. 2004; 3:958–964. [PubMed: 15473683]
- Gupta N, Pevzner PA. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*. 2009; 8:4173–4181. [PubMed: 19627159]

- Hawkrigde AM, Muddiman DC. Mass Spectrometry-Based Biomarker Discovery: Toward a Global Proteome Index of Individuality. *Annu Rev Anal Chem*. 2009; 2:265–277.
- Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biology*. 2003;4.
- Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*. 2007; 35:W169–W175. [PubMed: 17576678]
- Hughes C, Ma B, Lajoie GA. De novo sequencing methods in proteomics. *Methods Mol Biol*. 2010; 604:105–121. [PubMed: 20013367]
- Jaffe JD, Keshishian H, Chang B, Addona TA, Gillette MA, Carr SA. Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. *Molecular & cellular proteomics: MCP*. 2008; 7:1952–1962.
- Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA. PEPPer, a platform for experimental proteomic pattern recognition. *Molecular & cellular proteomics: MCP*. 2006; 5:1927–1941.
- Jiang X, Han G, Feng S, Ye M, Yao X, Zou H. Automatic validation of phosphopeptide identifications by the MS2/MS3 target-decoy search strategy. *Journal of proteome research*. 2008; 7:1640–1649. [PubMed: 18314942]
- Jung I, Matsuyama A, Yoshida M, Kim D. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC bioinformatics*. 2010; 11(Suppl 1):S10. [PubMed: 20122181]
- Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*. 2008; 7:29–34. [PubMed: 18067246]
- Kapp E, Schutz F, Simpson R. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms; Sensitivity and specificity analysis. *Molecular & Cellular Proteomics*. 2005; 4:S24–S24.
- Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular systems biology*. 2005; 1:2005 0017. [PubMed: 16729052]
- Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*. 2004; 20:3179–3184. [PubMed: 15231530]
- Krokhin OV, Ying S, Cortens JP, Ghosh D, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA. Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. *Analytical chemistry*. 2006; 78:6265–6269. [PubMed: 16944911]
- Lam H, Deutsch EW, Edes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 2007; 7:655–667. [PubMed: 17295354]
- Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology*. 2008; 4:222. [PubMed: 18854821]
- Leptos KC, Sarracino DA, Jaffe JD, Krastins B, Church GM. MapQuant: open-source software for large-scale protein quantification. *Proteomics*. 2006; 6:1770–1782. [PubMed: 16470651]
- Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd. Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology*. 1999; 17:676–682.
- Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry*. 2004; 76:4193–4201. [PubMed: 15253663]
- Lu B, Ruse C, Xu T, Park SK, Yates J 3rd. Automatic validation of phosphopeptide identifications from tandem mass spectra. *Analytical chemistry*. 2007; 79:1301–1310. [PubMed: 17297928]
- Lu B, Ruse CI, Yates JR 3rd. Colander: a probability-based support vector machine algorithm for automatic screening for CID spectra of phosphopeptides prior to database search. *Journal of proteome research*. 2008; 7:3628–3634. [PubMed: 18563924]

- Lundgren DH, Hwang SI, Wu L, Han DK. Role of spectral counting in quantitative proteomics. *Expert review of proteomics*. 2010; 7:39–53. [PubMed: 20121475]
- MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010; 26:966–968. [PubMed: 20147306]
- Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. eComputational prediction of proteotypic peptides for quantitative proteomics. *Nature biotechnology*. 2007; 25:125–131.
- Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105:18132–18138. [PubMed: 18818311]
- Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical chemistry*. 1994; 66:4390–4399. [PubMed: 7847635]
- Martin DB, Holzman T, May D, Peterson A, Eastham A, Eng J, McIntosh M. MRMer, an Interactive Open Source and Cross-platform System for Data Extraction and Visualization of Multiple Reaction Monitoring Experiments. *Molecular & Cellular Proteomics*. 2008; 7:2270–2278. [PubMed: 18641041]
- Mead JA, Bianco L, Ottone V, Barton C, Kay RG, Lilley KS, Bond NJ, Bessant C. MRmaid, the Web-based Tool for Designing Multiple Reaction Monitoring (MRM) Transitions. *Molecular & Cellular Proteomics*. 2009; 8:696–705. [PubMed: 19011259]
- Miller ML, Jensen LJ, Diella F, Jorgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovskiy M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R. Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*. 2008; 1:ra2. [PubMed: 18765831]
- Moore, RE.; Young, MK.; Lee, TD. Protein identification using a quadrupole ion trap mass spectrometer and SEQUEST database matching. In: Coligan, John E., et al., editors. *Current protocols in protein science*. Vol. Chapter 16. 2001. p. 10
- 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]
- Moruz L, Staes A, Foster JM, Hatzou M, Timmerman E, Martens L, Kall L. Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics*. 2012; 12:1151–1159. [PubMed: 22577017]
- Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Muller M. SuperHirn -a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*. 2007; 7:3470–3480. [PubMed: 17726677]
- Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*. 2010; 73:2092–2123. [PubMed: 20816881]
- Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*. 2003; 75:4646–4658. [PubMed: 14632076]
- Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*. 2007; 4:787–797. [PubMed: 17901868]
- Palagi PM, Walther D, Quadroni M, Catherinet S, Burgess J, Zimmermann-Ivol CG, Sanchez JC, Binz PA, Hochstrasser DF, Appel RD. MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics*. 2005; 5:2381–2384. [PubMed: 15880814]
- Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Analytical Chemistry*. 2009; 81:6481–6488. [PubMed: 19572557]
- Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology: CB*. 1993; 3:327–332. [PubMed: 15335725]
- Pedrioli PG. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol Biol*. 2010; 604:213–238. [PubMed: 20013374]
- Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb

- ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R. A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*. 2004; 22:1459–1466.
- 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]
- Picotti P, Lam H, Campbell D, Deutsch EW, Mirzaei H, Ranish J, Domon B, Aebersold R. A database of mass spectrometric assays for the yeast proteome. *Nature methods*. 2008; 5:913–914. [PubMed: 18974732]
- Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*. 2010; 11:395. [PubMed: 20650010]
- Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T. EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Molecular & cellular proteomics: MCP*. 2007; 6:527–536.
- Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*. 2009; 8:2405–2417. [PubMed: 19608599]
- Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, Hengartner MO, Aebersold R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature methods*. 2011; 8:430–435. [PubMed: 21423193]
- Rost H, Malmstrom L, Aebersold R. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Molecular & cellular proteomics: MCP*. 2012; 11:540–549.
- Ruttenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. *Journal of proteome research*. 2008; 7:3054–3059. [PubMed: 18543960]
- Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR 3rd. Code developments to improve the efficiency of automated MS/MS spectra interpretation. *Journal of proteome research*. 2002; 1:211–215. [PubMed: 12645897]
- Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, Kuster B. Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & cellular proteomics: MCP*. 2011; 10:M110 003830.
- Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. Proteomics-grade de novo sequencing approach. *Journal of proteome research*. 2005a; 4:2348–2354. [PubMed: 16335984]
- Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Molecular & cellular proteomics: MCP*. 2005b; 4:1180–1188.
- Schmidt A, Gehlenborg N, Bodenmiller B, Mueller LN, Campbell D, Mueller M, Aebersold R, Domon B. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Molecular & cellular proteomics: MCP*. 2008; 7:2138–2150.
- Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*. 2010; 10:1265–1269. [PubMed: 20077414]
- Sherwood CA, Eastham A, Lee LW, Peterson A, Eng JK, Shteynberg D, Mendoza L, Deutsch EW, Risler J, Tasman N, Aebersold R, Lam H, Martin DB. MaRiMba: A Software Application for Spectral Library-Based MRM Transition List Assembly. *Journal of Proteome Research*. 2009; 8:4396–4405. [PubMed: 19603829]
- Shifman MA, Li Y, Colangelo CM, Stone KL, Wu TL, Cheung KH, Miller PL, Williams KR. YPED: a web-accessible database system for protein expression analysis. *Journal of Proteome Research*. 2007; 6:4019–4024. [PubMed: 17867667]
- Stergachis AB, MacLean B, Lee K, Stamatoyannopoulos JA, MacCoss MJ. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nature Methods*. 2011; 8:1041–+. [PubMed: 22056677]

- Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O. OpenMS -an open-source software framework for mass spectrometry. *BMC Bioinformatics*. 2008; 9:163. [PubMed: 18366760]
- Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*. 2007; 6:654–661. [PubMed: 17269722]
- Tabb DL, Saraf A, Yates JR 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*. 2003; 75:6415–6421. [PubMed: 14640709]
- 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. *Analytical Chemistry*. 2005; 77:4626–4639. [PubMed: 16013882]
- Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, Henrich C, Mechtler K. Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research*. 2011; 10:5354–5362. [PubMed: 22073976]
- Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondrisk TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR 3rd, Hermjakob H. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*. 2007; 25:887–893.
- Turewicz M, Deutsch EW. Spectra, chromatograms, Metadata: mzML-the standard data format for mass spectrometer output. *Methods Mol Biol*. 2011; 696:179–203. [PubMed: 21063948]
- Vahamäa H, Koskinen VR, Hosia W, Moulder R, Nevalainen OS, Lahesmaa R, Aittokallio T, Salmi J. PolyAlign: A Versatile LC-MS Data Alignment Tool for Landmark-Selected and -Automated Use. *International journal of proteomics*. 2011; 2011:450290. [PubMed: 22084688]
- Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*. 2004; 1:39–45. [PubMed: 15782151]
- Vizcaino JA, Cote R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L. The Proteomics Identifications database: 2010 update. *Nucleic Acids Research*. 2010; 38:D736–742. [PubMed: 19906717]
- Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Research*. 2007; 35:W588–594. [PubMed: 17517770]
- Xu CX, Ma B. Software for computational peptide identification from MS-MS data. *Drug Discov Today*. 2006; 11:595–600. [PubMed: 16793527]
- Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual Review of Biomedical Engineering*. 2009; 11:49–79.
- Zerck A, Nordhoff E, Resemann A, Mirgorodskaya E, Suckau D, Reinert K, Lehrach H, Gobom J. An Iterative Strategy for Precursor Ion Selection for LC-MS/MS Based Shotgun Proteomics. *Journal of Proteome Research*. 2009; 8:3239–3251. [PubMed: 19402737]
- Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of proteome research*. 2006; 5:2909–2918. [PubMed: 17081042]
- Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics: MCP*. 2012; 11:M111 010587.
- Zhang N, Aebersold R, Schwikowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*. 2002; 2:1406–1412. [PubMed: 12422357]
- Zhang Y, Wen Z, Washburn MP, Florens L. Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Analytical Chemistry*. 2010; 82:2272–2281. [PubMed: 20166708]

- Zhou FF, Xue Y, Chen GL, Yao X. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochemical and biophysical research communications*. 2004; 325:1443–1448. [PubMed: 15555589]
- Zhu WH, Smith JW, Huang CM. Mass Spectrometry-Based Label-Free Quantitative Proteomics. *Journal of Biomedicine and Biotechnology*. 2010
- Zubarev RA, Zubarev AR, Savitski MM. Electron capture/transfer versus collisionally activated/induced dissociations: Solo or duet? *Journal of the American Society for Mass Spectrometry*. 2008; 19:753–761. [PubMed: 18499036]
- Zybailov B, Mosley AL, Sardi ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *Journal of Proteome Research*. 2006; 5:2339–2347. [PubMed: 16944946]