BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 5. 304–317
Advance Access publication June 20, 2007
doi:10.1093/bib/bbm023

Current trends in computational inference from mass spectrometry-based proteomics

Bobbie-Jo M. Webb-Robertson and William R. Cannon

Submitted: 31st March 2007; Received (in revised form): 23rd May 2007

Abstract

Mass spectrometry offers a high-throughput approach to quantifying the proteome associated with a biological sample and hence has become the primary approach of proteomic analyses. Computation is tightly coupled to this advanced technological platform as a required component of not only peptide and protein identification, but quantification and functional inference, such as protein modifications and interactions. Proteomics faces several key computational challenges such as identification of proteins and peptides from tandem mass spectra as well as their quantitation. In addition, the application of proteomics to systems biology requires understanding the functional proteome, including how the dynamics of the cell change in response to protein modifications and complex interactions between biomolecules. This review presents an overview of recently developed methods and their impact on these core computational challenges currently facing proteomics.

Keywords: *mass spectrometry; proteomics; peptide identification; post-translational modification; protein interaction*

INTRODUCTION

The explicit goal of proteomics is to identify and quantify all the proteins present in a cell at a specific moment. However, this is a significant challenge because unlike the genome, the proteins present in a system at any time are dynamic and of varying complexity. Multiple distinct proteins of differing function can arise from a single gene due to protein processing mechanisms such as alternative splicing or post-translational modifications (PTMs). The technique of choice for this formidable task has become mass spectrometry (MS). MS offers both a high-throughput (HTP) platform as well as the capability to process complex samples at a global scale. However, this HTP capability has led to a computational bottleneck as we struggle to analyze and interpret these large spectral datasets. In this briefing, we explore how the current state of research is tackling the computational challenges facing MS-based proteomics and the future prospects of global proteomics.

Understanding the computational needs and requirements of MS-based proteomics requires an underlying knowledge of the experimental process. Here, we describe a typical MS proteomics process from protein isolation through peptide identification; for detailed descriptions of the types of mass analyzers and specific details on characteristics and performance refer to Lane [1] and Domon and Aebersold [2]. The multi-scalar challenges of the MS-based proteomics process are hinted at in Figure 1—a typical MS-based proteomics analysis performed in many laboratories applicable to shotgun proteomics. As a first step, proteins are extracted from cells and

Corresponding author. Bobbie-Jo Webb-Robertson, Department of Computational Biology & Bioinformatics, Pacific Northwest National Laboratory, P.O. BOX 999, Richland, WA 99352, USA. Tel: 509-375-2292; Fax: 509-372-4720; E-mail: bj@pnl.gov

Bobbie-Jo M. Webb-Robertson is a research scientist in the Department of Computational Biology and Bioinformatics at Pacific Northwest National Laboratory. Her research interests include the application of statistical methods to bioinformatics problem such as protein inference, sequence homology and biomarker discovery.

William R. Cannon is a research scientist in the Department of Computational Biology and Bioinformatics at Pacific Northwest National Laboratory. His current interests include the development of statistical inference methods for high-throughput proteomics and network inference based on physical and biological principles.

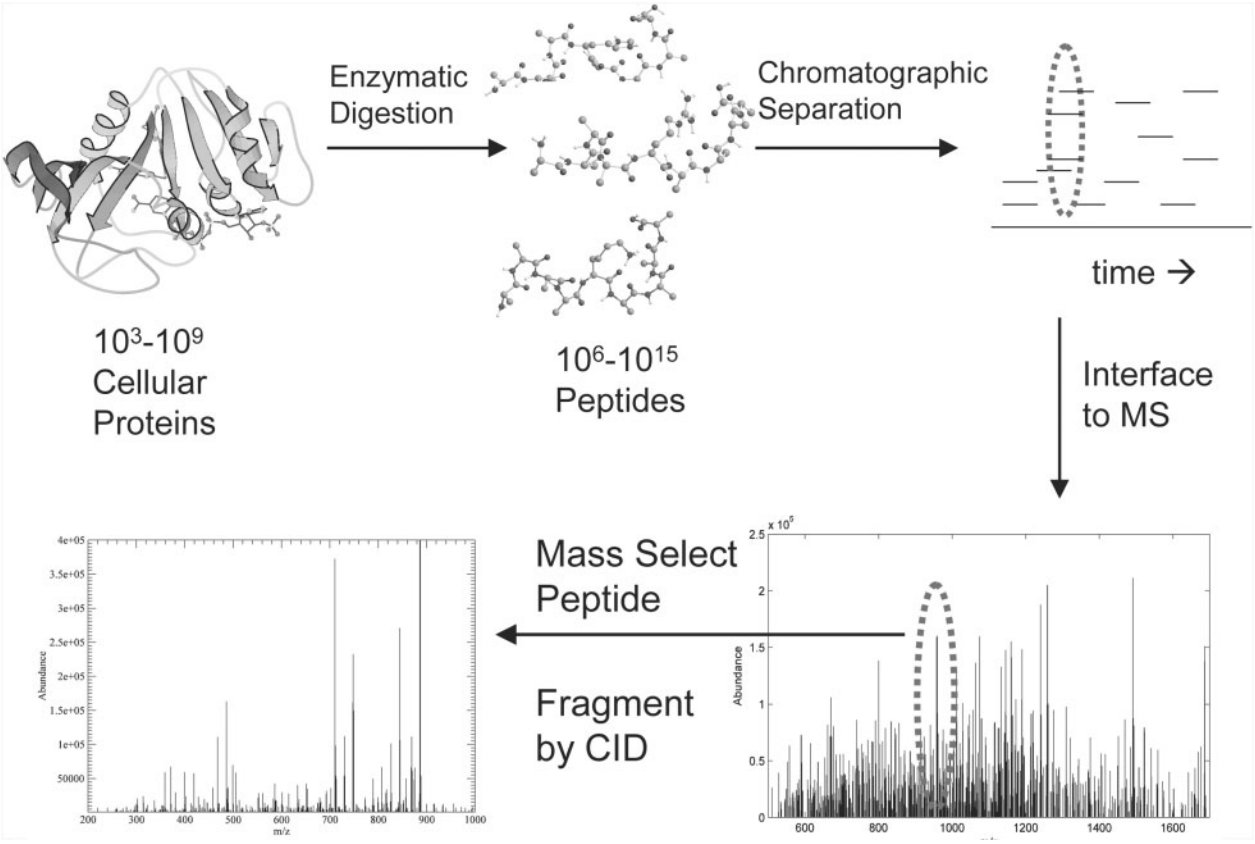

Figure 1: A typical MS proteomics process from protein isolation through peptide identification. Proteins undergo enzymatic digestion to fragment the proteins into peptides of size amenable to MS. The peptides are then separated via reverse phase high performance liquid chromatography and subjected to MS—an example separated by MS is shown by the ovals. A second phase of MS allows capture of the MS/MS spectrum of a single peptide.

cut at defined locations via enzymatic digestions. Single or multidimensional high performance liquid chromatography (HPLC) is then used to partially separate the peptides in the solution. The eluting peaks consist of populations of peptides, which are analyzed by a mass spectrometer interfaced with the HPLC system. The electrospray process nebulizes and ionizes the peptides into the gas phase and the charged particles are propelled into the mass spectrometer for analysis. The mass spectrometer scans the population of ions, measures the mass-to-charge ratio and proceeds to the second stage in the tandem process. This step consists of the capture of all ions in a narrow mass-to-charge range in an ion trap of the mass spectrometer, where the peptides are vibrationally excited by collision with inert gas. The peptides then fragment at labile bonds and a subsequent mass spectrum is obtained of the fragments of the peptide—a tandem mass spectrum (MS/MS), shown in the bottom panel of Figure 1.

Because peptides tend to fragment into recognizable patterns, the identity of peptides can frequently be determined from this MS/MS spectrum. The accuracy of this peptide identification step is critical to facilitate subsequent analyses. Additionally, contained in these spectra is information about the functional properties of peptides, such as PTMs. Recent years have seen a split in research between this peptide identification challenge and research based on these identifications to understand the dynamics of the system, such as quantitative differential expression and complex protein interactions.

THE PEPTIDE AND PROTEIN IDENTIFICATION CHALLENGE

There are two basic approaches that are taken for peptide identification from MS/MS data: database search [3–14] and *de novo* sequencing [15–23] represented in Figure 2. Database search methods

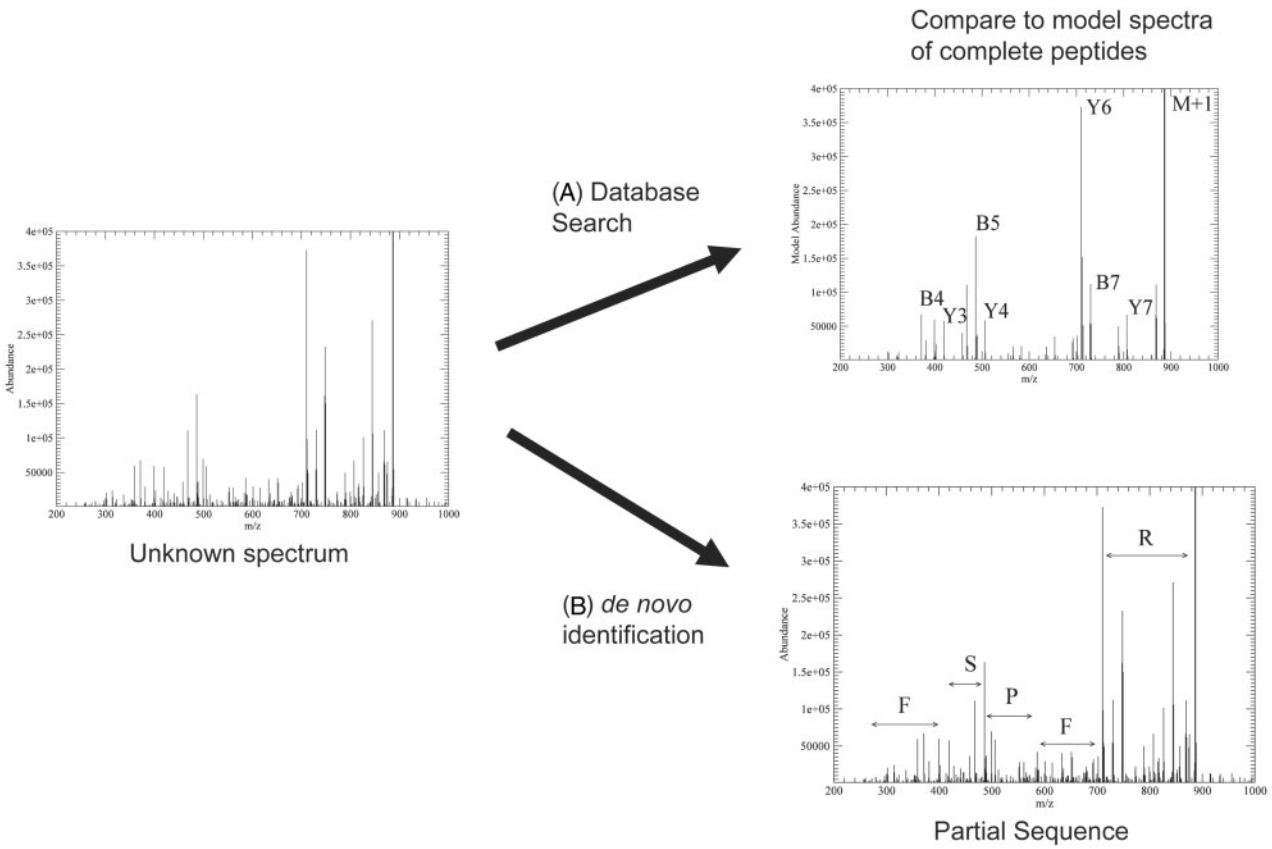

Figure 2: The two common approaches to peptide identifications. (**A**) Database search compares an acquired MS/MS spectrum to a database of model spectra derived from the genome of the organism under study. (**B**) *De novo* sequencing uses the mass differences between ion peaks to infer the peptide sequence directly from the acquired MS/MS spectrum.

have long been the primary method of peptide identification. This is largely because proposed *de novo* methods have traditionally required a minimal level of fragmentation and sequence coverage not realistic for collision induced dissociation (CID) spectra using available mass spectrometers. Both database search and *de novo* methods have seen a shift in the problems tackled in the literature the past few years. Database search methods are beginning to focus on more accurate identification, i.e. methods to address the large false positive problem introduced by searching large lists of peptides with generic fragmentation models and low information content spectra. *De novo* methods are focusing on the new mass spectrometers that have recently entered the market, such as the Linear Quadrupole ion Trap–Fourier Transform (LTQ–FT), Quadrupole-Time-of-Flight (Q-TOF) and Orbitrap, that offer ~2 orders of magnitude boost to the mass resolution compared to low-precision ion-trap detectors.

Reducing the false positives in database searches

Database search for peptide identification is the norm in proteomics. In this approach, a database of potential peptides are acquired from the genome of the organism under study. The sheer number of peptide candidates to consider for each spectrum, shown for various biological samples in Figure 3, leads to a potentially high number of false identifications. For each of the multitude of peptides, a model MS/MS spectrum is generated and compared against the experimental MS/MS spectrum based on a scoring metric. Out of this multitude of candidates, one peptide is identified following the scheme in Figure 2A. Due to the increasing need for accuracy as the protein databases are expanded, new algorithms for scoring an experimental to a model spectrum have largely focused on accounting for multiple sources of information derived from the experimental spectrum, such as mass peak intensity and

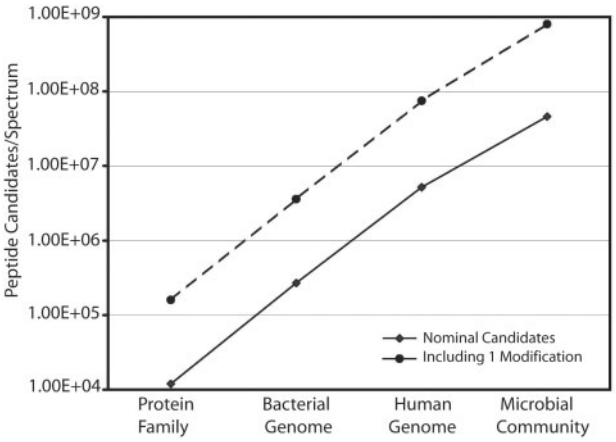

Figure 3: Average number of peptide candidates generated without regard to enzyme cleavage specificities that are to be considered for a match to each experimental spectrum. Using enzyme cleavage specificities can reduce the number of candidates by approximately one order of magnitude, while the inclusion of additional post-translational modifications can increase the candidates combinatorially. The sheer number of candidates can potentially lead to a large number of false identifications.

correlations among ions [24–26]. Beyond the mathematical form of the scoring metric, every scoring metric relies on two sources of information that are inherently important to accurate identification, (i) high quality MS/MS experimental spectra and (ii) properly represented model spectra. Better methods to represent these two key components have become a primary topic of interest in recent years.

It is clear that different technologies offer different resolutions and that the performance of different database searching algorithms can be dependent upon the platform the experiment was run under [27]. However, by understanding the limitations of different platforms, new methodologies can be developed that can manipulate these spectra to attain optimal performance. One primary consideration is to improve the quality of the identifications by developing methods to either filter spectra with insufficient information [28] or to remove low signal peaks (often referred to as noise, but which are often less common fragments) and hence improve the quality of the identifications [29, 30]. A second consideration is the peptides represented in the database. A commonly discussed approach to both reduce the raw size of the database and remove potential false positives is the generation of *proteotypic* peptides: the peptides that are detectable by a MS-experiment [31–34]. For example, peptides without ionizable amino acids carry only a minimal charge on the N terminus, reducing the likelihood of capturing that peptide by MS and thus the identification of that peptide in a database search. These approaches have clear relevance to the computational challenge of analyzing HTP proteomics data by offering a reduced search space and additionally are being demonstrated to be highly relevant beyond identification to quantitative proteomics [33, 34].

Beyond the peptides represented in the database, of utmost importance is generating realistic model spectra from these sequences in light of challenges such as partial fragmentation. Common models of spectra either represent the abundance of a series of peaks due to ion fragments as a set of constant values [7] or a statistical trained average probability of appearance [16]. A representative model spectrum for the constant and probability-based methods for des-Arg^9-bradykinin (PPGFSPFR) is shown in Figure 4B and 4C, respectively. Recent work has expanded on the statistical modeling idea by determining individual probabilities for each ion of an ion series [4], Figure 4D. Although, the latest fragmentation models are improved over generic models, no model spectra or fragmentation models are a close match to the actual spectrum shown for the peptide in Figure 4A. Recent work has focused on developing sequence-specific fragmentation models [35, 36]. This is important because actual fragmentation patterns due to CID vary tremendously based on the dynamical behavior of the peptide, which is dependent on the peptide sequence.

Given the challenge of predicting realistic model spectra an alternate approach has centered on the use of peptide MS/MS spectral libraries. Issues of sequence-dependent fragmentation can be modeled from these libraries [37–40], for example Breci *et al.* [37] characterizes the relative propensity of proline to fragment depending its N-terminal neighbor. These methods have tremendous potential for improving peptide identification; however, there are several problems with these approaches. First, given the fact that free energy contributions to fragmentation are not additive, the evaluation of the effects of pairs of amino acids is not sufficient for predicting the fragmentation patterns on longer peptides. Second, it is not possible to build a sufficiently diverse spectral database. For example, just for a 6-mer there

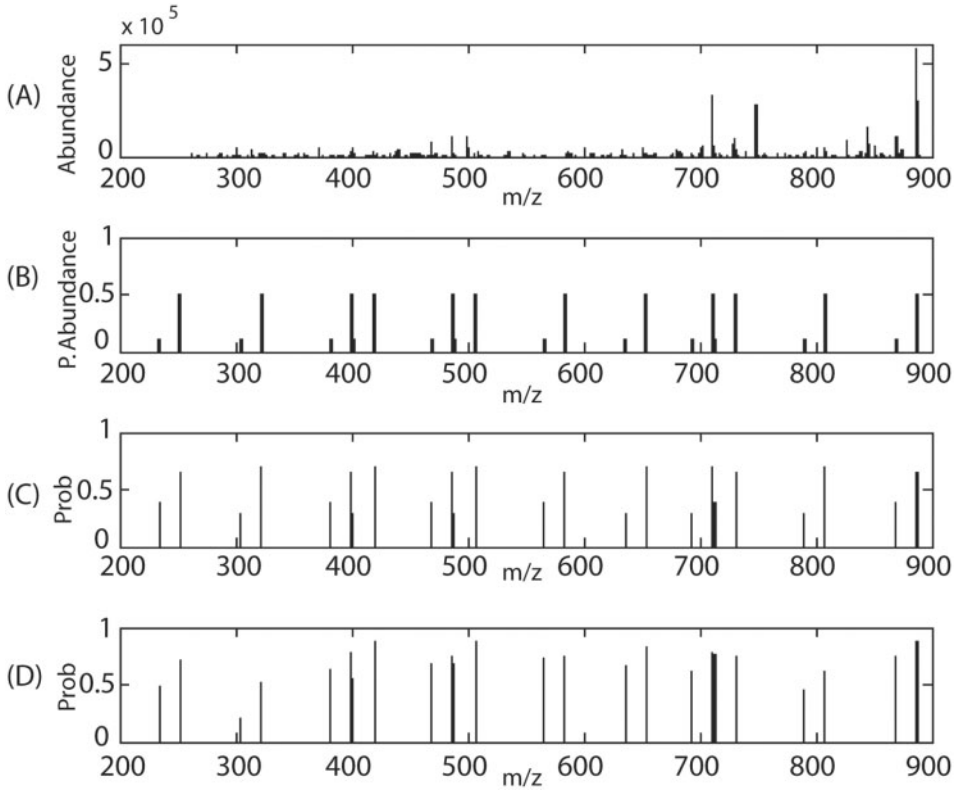

Figure 4: (**A**) Experimental MS/MS spectrum for des-Arg^9-bradykinin (PPGFSPFR). (**B**) Model spectrum similar to that used by SEQUEST in 1994. (**C**) Model spectrum similar to that used by Dancik *et al.* which was developed through the use of statistical training methods in which the probability of observing a peak at a specific location was learned from a sequence-averaged training set. (**D**) Model spectrum which expands on the model used by Dancik by allowing the probabilities to vary as a function of the location of the fragmenting bond.

are 20^6 unique peptides requiring replicate observations. Even if these 20^6 spectra were available, it would make more sense to use the spectra directly as model spectra.

An older approach to deal with insufficient sampling has gained significant momentum in the past few years—the idea of using of spectral libraries directly [31, 41–44]. In this approach, previously identified spectra for specific peptides serve as the model spectra to which newly acquired spectra are compared. Replicate spectra for a peptide can be used to reduce the total number of spectra to a set of consensus spectra that when combined are a more robust representation of the fragmentation patterns [45]. This adds increased confidence to the identification of newly acquired MS/MS spectra as long as the original spectra are correctly annotated. Similar to the identification of *proteotypic* peptides these approaches will likely have significant impact in the future of peptide identification as datasets continue to grow in size and more accurate identifications are needed for systems biology application. An important caveat to this approach, however, is that the original identifications rely on current peptide identification methods. Incorrect annotations will be propagated through the spectral libraries, similar to the phenomenon that has occurred in the gene annotation community. Additionally, the dependence on current peptide identification algorithms means that peptides that do not fragment completely are likely to be underrepresented in the database. Thus, while spectral libraries hold great promise for increasing peptide identifications, this approach must be combined with new methods for scoring peptides if the promise of spectral libraries is going to come to fruition.

Lastly, another approach that is commonly used to reduce false positive identifications from database searches is statistical measures of confidence. One of the best known applications of this approach is a statistical discriminant function generated for SEQUEST and peptide parameters in 2002, Peptide Prophet [46], and its protein counterpart

in 2003, Protein Prophet [47], although several approaches exist at the peptide [46, 48, 49] and protein [49–53] levels. Surprisingly, few new methods in this area have been proposed in the past few years, but the theme of these new papers focus on using decoy databases (known false candidates) to quantify false positive rates [49, 52]. These methods are more computationally costly, but more robust than their predecessors.

Higher accuracy *de novo* peptide sequencing

There are many possible cases for which a database search approach may not be able to identify a peptide associated with a MS/MS spectrum, such as novel proteins, mutations, PTMs, sequencing errors and community-based proteomics (meta-proteomics). In these cases, attaining the peptide directly from the MS/MS spectrum is desirable, which is the realm of *de novo* peptide sequencing [3, 15–23, 54]. In principle, a MS/MS spectrum contains a set of ions that can be used to compute the mass of the peptide and the distance between these ion peaks can be used to determine peptide composition, as shown in Figure 2B. However, in reality incomplete peptide fragmentation and low mass accuracy yield only partial sequence information. As a consequence, research on *de novo* algorithms seemed to lag in the late 1990s, but recent years have seen a new resurgence as new mass spectrometers that offer higher mass accuracy are coming on the market.

Methods that do not require statistical training may in principle more easily accommodate mechanisms for specifying mass accuracy levels in the search. Many *de novo* methods fall into this category. A few *de novo* methods such as SHERENGA [16] are exceptions in that they use statistical training. However, it may be the case that even reuse of derived information from training using differing mass accuracy parameters provides more accurate identifications than non-trained approaches. There are three primary approaches to performing *de novo* sequencing [55–58]. The first is graph theory which defines peaks as vertices and mass differences relate to edges. This is the traditional approach dating back to the start of *de novo* sequencing [55]. Graph methods are attractive because ideally they allow the user to identify the correct path simply by tracing a full pathway through the network [15, 16, 22]. Less effort has been placed on these methods compared to database searches in recent years [18], although Yan *et al.* [59] report an interesting application that uses graph theory to separate b and y ions in MS/MS spectra. Newer approaches for *de novo* sequencing are focusing on either optimization or sub-sequence matches augmented with sequence analysis.

Determining the correct sequence of amino acids for a peptide from MS/MS spectral data can be stated as a more general optimization problem where the objective is to match an experimental spectrum with the amino acid sequence most likely to produce it without necessarily having to revert to a graph-theoretical framework. Several optimization routines and fitness functions have been proposed [17, 18, 20]. Heredia-Langer *et al.* [20] propose a genetic algorithm that functions on spectral features that are often not captured in deterministic solution approaches. More recent work by Frank *et al.* [18] focus on the gain that can be attained from the high precision mass spectrometers FT-ICR, Q-TOF and Orbitrap. Likely pure *de novo* methods of the future will focus on these high mass accuracy machines.

Like the graph-theory methods, optimization approaches still have one primary caveat; they often return incomplete sequence information. Mann and Wilm initially addressed this problem well before proteomics became as high throughput as it is today by following a *de novo* identification with a search of the protein database using the peptide tag [60]. A recent surge in research has focused on coupling traditional *de novo* sequencing with sequence comparison methods [61–64]. One common approach, such as that by Wielsch *et al.* [65], use a tool known as MS BLAST [66] to match marginal quality MS/MS spectra. Although mass resolution is improving in the field of proteomics which increases the efficiency of pure *de novo* and optimization methods, it is likely that the incomplete nature of the information in MS/MS spectra will make sequence comparison a standard companion tool to complement *de novo* sequencing and optimization. Comparison of the discovered subsequences to protein databases also adds the benefit of extrapolating the peptide information to the protein level.

In the foreseeable future it is unlikely that *de novo* peptide sequencing will become the common approach to peptide identification. Database methods are generally more computationally friendly than *de novo* methods and offer a constrained space under which the protein identification problem is more easily tackled. However, *de novo* methods do potentially offer more flexibility. This capability

will become more important as samples become more complex, such as in environmental samples and meta-proteomics in which the relevant genomes are unknown or not fully sequenced [67].

PROTEIN QUANTIFICATION

The quantification of proteins in a sample has primarily been addressed using two approaches, isotopic labeling (ICAT, ITRAQ, SILAC, 18O- or 15N-labeling) approaches [63, 67–75] and label-free methods [76–78]. The primary limitation of proteomic labeling techniques is cost and a restriction on the number of conditions that can be compared based on the number of available isotopes. However, the demonstrated reproducibility of the method has continued to foster new computational methods development [68, 79, 80]. For example, Pan *et al.* [81, 82] have developed a profile likelihood algorithm that yields a point estimate and associated confidence for abundance that is more accurate than averages of peptide abundance ratios. As a result of the cost and added labor of the labeling approaches there has been a growing interest in label-free methods for quantification which are showing a surprising ability for reproducibility as well [76, 78].

Quantitative proteomics on isotopic labeled samples has a much longer history than nonlabeled methods [83, 84]. Traditional approaches have focused on the correlation between the datasets using statistical modeling methods such as regression. With the recent surge in higher resolution mass spectrometers and new labeling techniques, similar to *de novo* sequencing, new algorithms are being proposed that take advantage of these new technological advances [85–88]. For example Andreev *et al.* [85] use MS peak intensity measured from FT MS enhanced by a novel scoring algorithm based on 15N-labeling. Lin *et al.* [87] approaches focus on newer labeling methods such as iTRAQ using statistical models integrated with data-filtering thresholds.

The use of label-free proteomics for quantification is relatively new and hence the methods being developed in this area have a different flavor [34, 89–91]. Some of the most interesting work is related to understanding the underlying structure of the data. Callister *et al.* [90] address the issue of systematic bias related to relative quantification and identify linear regression normalization as one of the most robust normalization procedures to address this issue. Tang *et al.* [34] demonstrate that the concept of peptide detectability is highly relevant to quantification, specifically that there is a correlation between the quantity of a protein and the likelihood of its constituent peptides to be identified by MS. In addition, in place of direct measurements of peptide or protein abundance, attempts have been made to use proxies such as the number of spectra that can be assigned to a given peptide [92, 93]. Methods that allow for the direct quantitation will be important to the future of label-free proteomics, as well as being highly relevant to the isotopic labeling strategies.

THE FUNCTIONAL PROTEOME—POST-TRANSLATIONAL MODIFICATIONS

Modified proteins are abundant in living systems and are known to have profound biological implications. The covalent modification of amino acid residues that are the result of a PTM fit naturally with MS because the cleavage fragment can be identified by characteristic mass shifts. For example, one of the most common PTMs, phosphorylation on serine or threonine, yields a mass increase of 80 units in the molecular weights of the intact unmodified peptide and its expected fragments and a 98 and 49 neutral mass unit losses from key peaks in the MS/MS spectrum of the modified peptide. However, PTMs are numerous, complex and typically involve either proteolytic cleavage or the addition of a modifying group to one or more amino acids. Despite the mass-modifying properties of PTMs that make them amenable to identification by MS, both the shear number and the biological complexity [94, 95] continue to make PTM identification a challenge. A large amount of effort is required to build an accurate model of a single PTM, as evidenced by many review articles regarding the MS of phosphorylation [96–99].

The most obvious strategy for PTM is known as spectral alignment and was first introduced by Yates *et al.* in 1995 [100]: augment the search database of model spectra to contain all common modifications. The combinatorics of evaluating all modifications quickly grows to a computationally intractable problem for most standard MS laboratories. Subsequent work focused on reducing this search space by deriving potential peptides for modification based on a first pass protein identification step [101–104]. This can still be highly computationally

challenging and is dependent upon the accuracy of the first step of protein identification. Recent research involving spectral alignment have focused on removing the *known* modifications component from the search and improved scoring methods. Hansen *et al*. [105] search identified peptides against the experimental spectra to compute mass differences. They use these mass differences, in conjunction with a statistical model, to identify the localization of the modification. In a more generalized manner Tsur *et al*. [106] demonstrate a local alignment procedure that increases comparative speed and does not require preexisting knowledge of PTMs. A frequency histogram of mass differences is used to determine present modifications. Alternatively, Havilio and Wool [107] use a fast-Fourier transform to look for mass offsets in a rapid, unrestricted manner. These approaches are very promising as they tie in nicely with database searching algorithms.

An alternative approach to alleviate this computational problem is to match the MS/MS spectra to the database without a parent ion mass tolerance [104, 108]. The problem with this approach traditionally is that such leniency in the matching step creates a significant false positive problem. The most recent work in this area uses *de novo* sequencing to build a target peptide sequence without reference to a database [109–111]. For example, the most current OpenSea approach of Searle *et al*. [110] uses a heuristic branch-and-bound technique to compare the *de novo* prediction to identify homologous sequences from a database. Another algorithm, SPIDER of Han *et al*. [111], compares the *de novo* sequence to the protein database. In both cases, these algorithms are highly dependent on accurate *de novo* interpretations. As *de novo* predictions improve these methods will become more attractive.

Only in the past few years have attempts been taken to integrate these two methods [112–114]. The MOD^i approach developed by Kim *et al*. [113] includes peak selection, tag discovery (*de novo*), database search, tag chain extension and finally PTM identification and is available as a web server. Overall the proteomics community has only begun to address this challenge of identification of PTMs in a routine and robust manner. These new methods are a step in the right direction, more computationally friendly and integrating more sources of information. This area of research is likely to remain a hot topic in the years to come. Reaching the goal of identification of all PTMs in a sample will be essential to map the proteome to meaningful function.

COMPLEX SYSTEM DYNAMICS—PROTEIN INTERACTIONS

MS-based proteomics analyses follow one of two paradigms, shotgun proteomics in which the goal is to analyze the entire expressed proteome of a cell population at a time [115, 116], or targeted proteomics in which the goal is to cover a specific protein-related phenomena in depth. The latter has a long history starting with the first applications of mass spectrometry to proteins. In contrast, shotgun proteomics has been a recent and fastly growing approach that has been catalyzed by the combination of the rapid pace of genome sequencing, breakthroughs in mass spectrometry technology development and the application of large-scale computing for data analysis. In this section, we briefly discuss some of the issues regarding the targeted proteomics analysis of protein–protein interactions, which is focused both on specific protein interactions yet is relatively high-throughput [117–119].

One of the greatest challenges in analyzing high-throughput protein interaction data is the determination of which interactions are present because they are true *in vivo* interactions, and which are present as an artifact of the experimental protocol. Artifactual interactions are present for a number of reasons. First, the process of lysing the cell destroys the compartmentalization that is in part responsible for the specificity of biological interactions. A common assumption in large scale tandem affinity purifications is that the protein complex survives the experimental process intact [117, 118]. This assumption depends not only on the severity of the assay conditions, but also on the thermodynamic half-life of the complex itself. This problem is complicated by differences in the technical details of individual isolation procedures and by the fact that not all proteins are equally detected by each experimental protocol. A typical approach taken to eliminate nonspecific interactors is to remove prey proteins that show up in more than a few percent, typically ~3%, of the affinity assays. This approach makes sense but ultimately needs to be implemented within an objective statistical procedure; otherwise the results are obtained in an *ad hoc* manner that may not be reproducible.

Toward this goal, several different statistical approaches have recently been introduced that attempt to derive a confidence statistic for the likelihood of a given protein interaction pair being 'real'. In previous years these approaches were focused on assessing a variety of independent data with the idea that nonspecific or noisy interactions would drop out. Gerstein and colleagues [120, 121] have shown the value of combining data from multiple isolation approaches to derive an overall confidence value for the assignment of a protein to a complex. Gilchrist *et al*. [122] have developed a statistical framework for combining different types of proteomics data based on the assumption that nonspecific interactions are likely to be technology-specific (i.e. dependent on the isolation technique being used). Bader *et al*. [123] have shown that it is possible to define a quantitative confidence measure based entirely on screening statistics and network topology.

However, recent research suggests that instead of combining data from multiple low confidence sources it is possible to implement more focused statistical analysis in addition to advanced experimental designs that provide reliable information on protein interactions. For years protein interactions have been carefully characterized by small scale studies using technologies such as surface plasmon resonance and fluorescence resonance to obtain information on binding affinities. It may be that some of the concepts used in these small-scale studies can be scaled up for use in high-throughput proteomics. For example, Rinner *et al*. [124] recently reported the use of dilution studies to provide enough dynamic range in order to differentiate nonspecific interactions from specific interaction. Sufficient dynamic range is important because nonspecific interactions typically result from high abundance of the prey protein in solution compared to specific interactions, which tend to have lower prey abundance in solution and lower dissociation constants. Toward this goal, Sharp *et al*. [125] have recently developed a statistical assessment of protein interactions based on replicate measurements that can in principle discriminate between specific and nonspecific interactions. The combined use of more sophisticated experimental designs and statistical analyses should lead to the confident determination of protein interactions without the need to combine multiple sources of data.

Ideally, we would like to know what the essential proteins are that encode a given biological capability. However, at this point in our current understanding of complex systems this question can only be answered experimentally. Mathematical analysis of protein interaction networks alone cannot define protein complexes or signaling networks unless these analyses are coupled with specific follow-on functional experiments. For example, Scholten *et al*. [126] present a thoughtful local model analysis of protein interactions in which they use maximally complete subgraphs to estimate a complex. A maximally complete subgraph is defined as a graph which is the maximal graph that includes all pair-wise interactions. However, it is likely that the full complement of proteins required to encode many biological capabilities will not be represented by maximally complete subgraphs. For example, in the T4 phage DNA polymerase system the protein complex consists of the core polymerase, a clamp protein, and accessory loading proteins, which can assemble onto DNA through multiple kinetic pathways [127]. However, the thermodynamics of the pair-wise protein–protein interactions are not simply additive, with the result being that a preferred assembly order may exist [128]. This would imply that the biological capability encoded by this system may not be a maximally complete subgraph as determined by affinity purifications. However, it may be possible through the use of bioinformatics to relieve the strong assumption of maximally complete subgraphs.

Starting from the acknowledgment that mathematics alone cannot currently define a *biological* complex, we can set the more modest goal of using mathematically defined entities to further understand, perhaps decompose, a biological capability. In addition, each mathematically defined entity can be used to formulate the hypothesis to be tested experimentally that the entity represents a biological capability. This is consistent with the conclusion reached by Scholten *et al*. [126], 'Local modeling provides a platform for explicit hypothesis development regarding functional annotation and pathway activity.' Toward this goal, then we should ultimately seek to link biological experiments to network analysis of protein interactions in a biologically-principled manner.

PROTEOMICS STANDARDS

Proteomics standards is an interesting topic in proteomics because it really requires a commitment

from both the experimental and computational communities. The need for guidelines in supplementary information that should accompany results remains a debate; however, no researcher will debate that a minimum of information to reproduce the results is necessary. In regards to experimental data an editorial by Carr *et al*. [129] nicely lays out the challenges in implementing such requirements. A working group of the Proteomics Standards Initiative (PSI) [130] of the Human Proteome Organization have been working on a platform called MIAPE [131, 132] to capture proteomics experimental meta-data. In a similar vein as the microarray MIAME platform, MIAPE requires that the minimum information necessary to reproduce a proteomics experiment is captured. Adoption is slow but MIAPE, or something similar, will likely become a requirement of journals in the near future. On the computational side, the Institute of Systems Biology pioneered a generic XML representation of MS data called mzXML [133, 134]. On a similar track PSI has also been developing a generic XML based language for representing MS data and the results of computational analyses called mzData and analysisXML [132]. A recent article by Cottingham [135] states that these two standards are currently working on merging. Whatever the final standards will be, the need is becoming pressing and a foreseeable adoption of a single standard will happen as it is likely that journals and funding agencies will make these compliances a requirement.

CONCLUSIONS AND FUTURE DIRECTIONS

The next few years will be an exciting time in proteomics. As researchers are realizing the systems level biological information that HTP proteomics offers the computational challenges for proteomics will continue to grow. The data sets are becoming larger either due to sample size or available mass analyzers, generating easily tens to hundreds of thousands or more spectra; the computational load on peptide identification and downstream biological inferences, such as protein interactions becomes restrictive. An even larger challenge has presented itself recently as Lo *et al.* [136] demonstrated the biological relevance of proteomics on community samples. As more complex samples are analyzed the resources required for analysis are enormous as there is a potential for millions of proteins.

With this in mind, there will likely be two trends in MS-based proteomics, (i) system level analyses of well characterized organisms and (ii) meta-proteomics of complex and large samples with little annotation. For the first scenario it will be ultimately important to improve model spectra for more accurate peptide identification and characterization of PTMs. This will undoubtedly improve any downstream inferences of complex interactions between proteins. The second scenario of meta-proteomics is likely to take off in a similar manner as meta-genomics [137–141]. The simple analysis task will likely require either significant reduced database representations, such as those proposed by proteotypic peptides or spectral libraries, or enhanced sequence analysis *de novo* sequencing strategies that better handle partial fragmentation. Searching this data to ultimately identify community-based metabolic activity would offer new and exciting biological information to the world of ecology. All in all, this is an extremely exciting time in proteomics as the limits of the technology are still unknown and so much is still left to be discovered.

Key Points

- The proteins present in a system at any time are dynamic and of unknown complexity.
- Database searching offers a constrained space under which the protein identification problem is more easily tackled.
- *De novo* sequencing offers significant flexibility for peptide identification.
- New approaches to PTM identification that integrate spectral alignment with *de novo* sequence comparative methods are more computationally friendly and potentially more robust.
- It is possible to implement more focused statistical analysis in addition to advanced experimental designs that provide reliable information on protein interactions.

Acknowledgements

This work was supported through the US Department of Energy (DOE) Office of Advanced Scientific Computing Research under Contract No. 47901. PNNL is a multiprogram national laboratory operated by Battelle for the US DOE under contract DE-AC06-76L01830.

References

1. Lane CS. Mass spectrometry-based proteomics in the life sciences. *Cell Mol Life Sci* 2005;**62**:848–69.
2. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science* 2006;**312**:212–7.

3. Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001;**17**(Suppl 1):S13–21.
4. Cannon WR, Jarman KH, Webb-Robertson BJ, *et al*. Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data. *J Proteome Res* 2005;**4**:1687–98.
5. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**:1466–7.
6. Elias JE, Gibbons FD, King OD, *et al*. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol* 2004;**22**:214–9.
7. 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* 1994;**5**:976–89.
8. Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem* 2003;**75**:435–44.
9. Li D, Fu Y, Sun R, *et al*. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* 2005;**21**:3049–50.
10. Perkins DN, Pappin DJ, Creasy DM, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999; **20**:3551–67.
11. Sadygov RG, Yates JR, 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem* 2003;**75**:3792–8.
12. 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**:1406–12.
13. Geer LY, Markey SP, Kowalak JA, *et al*. Open mass spectrometry search algorithm. *J Proteome Res* 2004; **3**:958–64.
14. Liu J, Ma B, Li M. PRIMA: peptide robust identification from MS/MS spectra. *J Bioinform Comput Biol* 2006; **4**:125–38.
15. Chen T, Kao MY, Tepel M, *et al*. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001;**8**:325–37.
16. Dancik V, Addona TA, Clauser KR, *et al*. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999;**6**:327–42.
17. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005; **77**:964–73.
18. Frank AM, Savitski MM, Nielsen ML, *et al*. De novo peptide sequencing and identification with precision mass spectrometry. *J Proteome Res* 2007;**6**:114–23.
19. Halligan BD, Ruotti V, Twigger SN, *et al*. DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from de novo peptide sequencing by mass spectroscopy. *Nucleic Acids Res* 2005;**33**:W376–81.
20. Heredia-Langner A, Cannon WR, Jarman KD, *et al*. Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics* 2004; **20**:2296–304.
21. Savitski MM, Nielsen ML, Kjeldsen F, *et al*. Proteomics-grade de novo sequencing approach. *J Proteome Res* 2005; **4**:2348–54.
22. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997;**11**:1067–75.
23. Taylor JA, Johnson RS. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* 2001;**73**:2594–604.
24. MacLean B, Eng JK, Beavis RC, *et al*. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* 2006; **22**:2830–2.
25. Sun S, Meyer-Arendt K, Eichelberger B, *et al*. Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol Cell Proteomics* 2007; **6**:1–17.
26. Wan Y, Yang A, Chen T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *Anal Chem* 2006;**78**:432–7.
27. Elias JE, Haas W, Faherty BK, *et al*. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* 2005;**2**:667–75.
28. Flikka K, Martens L, Vandekerckhove J, *et al*. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* 2006; **6**:2086–94.
29. Bensmail H, Golek J, Moody MM, *et al*. A novel approach for clustering proteomics data using Bayesian fast Fourier transform. *Bioinformatics* 2005;**21**:2210–24.
30. Mujezinovic N, Raidl G, Hutchins JR, *et al*. Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics* 2006;**6**:5117–31.
31. Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom* 2005;**19**:1844–50.
32. Kuster B, Schirle M, Mallick P, *et al*. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* 2005;**6**:577–83.
33. Mallick P, Schirle M, Chen SS, *et al*. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 2007;**25**:125–31.
34. Tang H, Arnold RJ, Alves P, *et al*. A computational approach toward label-free protein quantification using predicted peptide detectability, *Bioinformatics* 2006; **22**:e481–8.
35. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem* 2005;**77**:6364–73.
36. Arnold RJ, Jayasankar N, Aggarwal D, *et al*. A machine learning approach to predicting peptide fragmentation spectra. *Pac Symp Biocomput* 2006;**11**:219–30.
37. Breci LA, Tabb DL, Yates JR, 3rd, *et al*. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal Chem* 2003;**75**:1963–71.
38. Huang Y, Wysocki VH, Tabb DL, *et al*. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int J Mass Spectrom* 2002;**219**:233–44.

39. Kapp EA, Schutz F, Reid GE, *et al*. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* 2003; **75**:6251–64.
40. Tabb DL, Smith LL, Breci LA, *et al*. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* 2003;**75**:1155–63.
41. Aebersold R. Constellations in a cellular universe. *Nature* 2003;**422**:115–6.
42. Beer I, Barnea E, Ziv T, *et al*. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004;**4**:950–60.
43. Craig R, Cortens JC, Fenyo D, *et al*. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* 2006;**5**:1843–9.
44. Yates JR, 3rd, Morgan SF, Gatlin CL, *et al*. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal Chem* 1998;**70**:3557–65.
45. Lam H, Deutsch EW, Eddes JS, *et al*. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007; **7**:655–67.
46. Keller A, Nesvizhskii AI, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;**74**:5383–92.
47. Nesvizhskii AI, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;**75**:4646–58.
48. Anderson DC, Li W, Payan DG, *et al*. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res* 2003; **2**:137–46.
49. Huttlin EL, Hegeman AD, Harms AC, *et al*. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J Proteome Res* 2007;**6**:392–8.
50. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 2003; **75**:768–74.
51. Ganapathy A, Wan XF, Wan J, *et al*. Statistical assessment for mass-spec protein identification using peptide fingerprinting approach. *Conf Proc IEEE Eng Med Biol Soc* 2004; **4**:3051–4.
52. Higdon R, Kolker E. A predictive model for identifying proteins by a single peptide match. *Bioinformatics* 2007; **23**:277–80.
53. Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom* 2002;**13**:378–86.
54. Bartels C. Fast algorithm for peptide sequencing by mass spectrometry. *Biomed Environ Mass Spectrom* 1990;**19**:363–8.
55. Pevtsov S, Fedulova I, Mirzaei H, *et al*. Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res* 2006;**5**:3018–28.
56. Johnson RS, Davis MT, Taylor JA, *et al*. Informatics for protein identification by mass spectrometry. *Methods* 2005; **35**:223–36.
57. Reinders J, Lewandrowski U, Moebius J, *et al*. Challenges in mass spectrometry-based proteomics. *Proteomics* 2004; **4**:3686–703.
58. Shadforth I, Crowther D, Bessant C. Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics* 2005; **5**:4082–95.
59. Yan B, Pan C, Olman VN, *et al*. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* 2005;**21**:563–74.
60. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;**66**:4390–9.
61. Frank A, Tanner S, Bafna V, *et al*. Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res* 2005;**4**:1287–95.
62. Habermann B, Oegema J, Sunyaev S, *et al*. The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol Cell Proteomics* 2004;**3**:238–49.
63. Han DK, Eng J, Zhou H, *et al*. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 2001;**19**:946–51.
64. Tabb DL, Saraf A, Yates JR, 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003;**75**:6415–21.
65. Wielsch N, Thomas H, Surendranath V, *et al*. Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches. *J Proteome Res* 2006;**5**:2448–56.
66. Shevchenko A, Sunyaev S, Loboda A, *et al*. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* 2001; **73**:1917–26.
67. Banfield JF, Verberkmoes NC, Hettich RL, *et al*. Proteogenomic approaches for the molecular characterization of natural microbial communities. *Omics* 2005; **9**:301–33.
68. Aggarwal K, Choe LH, Lee KH. Shotgun proteomics using the iTRAQ isobaric tags. *Brief Funct Genomic Proteomic* 2006;**5**:112–20.
69. Desiderio DM, Kai M. Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biomed Mass Spectrom* 1983;**10**:471–9.
70. Griffin TJ, Han DK, Gygi SP, *et al*. Toward a high-throughput approach to quantitative proteomic analysis: expression-dependent protein identification by mass spectrometry. *J Am Soc Mass Spectrom* 2001; **12**:1238–46.
71. Gygi SP, Rist B, Gerber SA, *et al*. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;**17**:994–9.
72. Hardt M, Witkowska HE, Webb S, *et al*. Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal Chem* 2005; **77**:4947–54.
73. Ong SE, Blagoev B, Kratchmarova I, *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple

and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002;**1**:376–86.

74. Patton WF. Detection technologies in proteome analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* 2002;**771**:3–31.
75. Sakai J, Kojima S, Yanagi K, *et al*. 18O-labeling quantitative proteomics using an ion trap mass spectrometer. *Proteomics* 2005;**5**:16–23.
76. Smith RD, Shen Y, Tang K. Ultrasensitive and quantitative analyses from combined separations-mass spectrometry for the characterization of proteomes. *Acc Chem Res* 2004; **37**:269–78.
77. Voyksner RD, Lee H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. *Rapid Commun Mass Spectrom* 1999; **13**:1427–37.
78. Wang G, Wu WW, Zeng W, *et al*. Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J Proteome Res* 2006; **5**:1214–23.
79. Chong PK, Gan CS, Pham TK, *et al*. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *J Proteome Res* 2006; **5**:1232–40.
80. Molloy MP, Donohoe S, Brzezinski EE, *et al*. Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling. *Proteomics* 2005; **5**:1204–8.
81. Pan C, Kora G, McDonald WH, *et al*. ProRata: a quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Anal Chem* 2006;**78**:7121–31.
82. Pan C, Kora G, Tabb DL, *et al*. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal Chem* 2006;**78**:7110–20.
83. MacCoss MJ, Wu CC, Liu H, *et al*. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* 2003;**75**:6912–21.
84. von Haller PD, Yi E, Donohoe S, *et al*. The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. *Mol Cell Proteomics* 2003;**2**:428–42.
85. Andreev VP, Li L, Rejtar T, *et al*. New algorithm for 15N/14N quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. *J Proteome Res* 2006;**5**:2039–45.
86. Faca V, Coram M, Phanstiel D, *et al*. Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *J Proteome Res* 2006;**5**:2009–18.
87. Lin WT, Hung WN, Yian YH, *et al*. Multi-Q: a fully automated tool for multiplexed protein quantitation. *J Proteome Res* 2006;**5**:2328–38.
88. Snijders AP, de Vos MG, Wright PC. Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling. *J Proteome Res* 2005;**4**:578–85.
89. Johansson C, Samskog J, Sundstrom L, *et al*. Differential expression analysis of *Escherichia coli* proteins using a novel software for relative quantitation of LC-MS/MS data. *Proteomics* 2006;**6**:4475–85.
90. Callister SJ, Barry RC, Adkins JN, *et al*. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 2006;**5**:277–86.
91. Fischer B, Grossmann J, Roth V, *et al*. Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* 2006;**22**:e132–40.
92. 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**:4193–201.
93. Zhang B, VerBerkmoes NC, Langston MA, *et al*. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 2006;**5**:2909–18.
94. Graves DJ, Martin BL, Wang JH. *Co- and Post-translational Modification of Proteins: Chemical Principles and Biological Effects*. New York: Oxford University Press, 1994.
95. Krishna RG, Wold F. *Posttranslational Modifications*. San Diego: Academic Press, 1998.
96. Areces LB, Matafora V, Bachi A. Analysis of protein phosphorylation by mass spectrometry. *Eur J Mass Spectrom* (Chichester, Eng) 2004;**10**:383–92.
97. Goshe MB. Characterizing phosphoproteins and phosphoproteomes using mass spectrometry. *Brief Funct Genomic Proteomic* 2006;**4**:363–76.
98. Mann M, Ong SE, Gronborg M, *et al*. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol* 2002;**20**:261–8.
99. Peters EC, Brock A, Ficarro SB. Exploring the phosphoproteome with mass spectrometry. *Mini Rev Med Chem* 2004;**4**:313–24.
100. Yates JR, 3rd, Eng JK, McCormack AL, *et al*. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 1995;**67**:1426–36.
101. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2002;**2**:1426–34.
102. Gatlin CL, Eng JK, Cross ST, *et al*. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal Chem* 2000;**72**:757–63.
103. MacCoss MJ, McDonald WH, Saraf A, *et al*. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci USA* 2002; **99**:7900–5.
104. Pevzner PA, Mulyukov Z, Dancik V, *et al*. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res* 2001;**11**:290–9.
105. Hansen BT, Davey SW, Ham AJ, *et al*. P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. *J Proteome Res* 2005; **4**:358–68.
106. Tsur D, Tanner S, Zandi E, *et al*. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 2005;**23**:1562–7.
107. Havilio M, Wool A. Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. *Anal Chem* 2007;**79**:1362–8.

108. Clauser KR, Baker P, Burlingame AL. 'Peptide fragment-ion tags from MALDI/PSD for error-tolerant searching of genomic databases', *44th ASMS Conference on Mass Spectrometry and Allied Topics, 1996*. Portland, OR.
109. Searle BC, Dasari S, Turner M, *et al*. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal Chem* 2004; **76**:2220–30.
110. Searle BC, Dasari S, Wilmarth PA, *et al*. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *J Proteome Res* 2005; **4**:546–54.
111. 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.
112. Tanner S, Shu H, Frank A, *et al*. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 2005;**77**:4626–39.
113. Kim S, Na S, Sim JW, *et al*. MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. *Nucleic Acids Res* 2006;**34**:W258–63.
114. Matthiesen R, Bunkenborg J, Stensballe A, *et al*. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* 2004;**4**:2583–93.
115. Wolters DA, Washburn MP, Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001;**73**:5683–90.
116. Smith RD, Anderson GA, Lipton MS, *et al*. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002; **2**:513–23.
117. Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;**415**:180–3.
118. Gavin AC, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**:141–7.
119. Verma R, Chen S, Feldman R, *et al*. Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol Biol Cell* 2000;**11**:3425–39.
120. Lin N, Wu B, Jansen R, *et al*. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics* 2004;**5**:154.
121. Lu LJ, Xia Y, Paccanaro A, *et al*. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* 2005;**15**:945–53.
122. Gilchrist MA, Salter LA, Wagner A. A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics* 2004;**20**:689–700.
123. Bader JS, Chaudhuri A, Rothberg JM, *et al*. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 2004;**22**:78–85.
124. Rinner O, Mueller LN, Hubalek M, *et al*. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat Biotechnol* 2007;**25**:345–52.
125. Sharp JL, Anderson KK, Daly DS, *et al*. Inferring protein associations using protein pull-down assays. *Proc Am Stat Ass* 2006.
126. Scholtens D, Vidal M, Gentleman R. Local modeling of global interactome networks. *Bioinformatics* 2005; **21**:3548–57.
127. Smiley RD, Zhuang Z, Benkovic SJ, *et al*. Single-molecule investigation of the T4 bacteriophage DNA polymerase holoenzyme: multiple pathways of holoenzyme formation. *Biochemistry* 2006;**45**:7990–7.
128. Sexton DJ, Kaboord BF, Berdis AJ, *et al*. Dissecting the order of bacteriophage T4 DNA polymerase holoenzyme assembly. *Biochemistry* 1998;**37**:7749–56.
129. Carr S, Aebersold R, Baldwin M, *et al*. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* 2004;**3**:531–3.
130. Orchard S, Hermjakob H, Apweiler R. The proteomics standards initiative. *Proteomics* 2003;**3**:1374–6.
131. Orchard S, Jones AR, Stephan C, *et al*. The HUPO Pre-Congress Proteomics Standards Initiative Workshop HUPO 5(th) Annual World Congress Long Beach, CA, USA 28 October-1 November 2006, Proteomics 2007.
132. Orchard S, Taylor CF, Hermjakob H, *et al*. Advances in the development of common interchange standards for proteomic data. *Proteomics* 2004;**4**:2363–5.
133. Lin SM, Zhu L, Winter AQ, *et al*. What is mzXML good for? *Expert Rev Proteomics* 2005;**2**:839–45.
134. Pedrioli PG, Eng JK, Hubley R, *et al*. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004; **22**:1459–66.
135. Cottingham K. Toward a single MS format. *J Proteome Res* 2006;**5**:15.
136. Lo I, Denef VJ, Verberkmoes NC, *et al*. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 2007.
137. Cowan D, Meyer Q, Stafford W, *et al*. Metagenomic gene discovery: past, present and future. *Trends Biotechnol* 2005; **23**:321–329.
138. Handelsman J, Rondon MR, Brady SF, *et al*. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998;**5**:R245–9.
139. Markowitz VM, Ivanova N, Palaniappan K, *et al*. An experimental metagenome data management and analysis system. *Bioinformatics* 2006;**22**:e359–67.
140. Rondon MR, August PR, Bettermann AD, *et al*. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* 2000; **66**:2541–7.
141. Streit WR, Schmitz RA. Metagenomics—the key to the uncultured microbes. *Curr Opin Microbiol* 2004; **7**:492–8.