Analysis and validation of proteomic data generated by tandem mass spectrometry

Alexey I Nesvizhskii[1], Olga Vitek[2] & Ruedi Aebersold[3,4]

The analysis of the large amount of data generated in mass spectrometry–based proteomics experiments represents a significant challenge and is currently a bottleneck in many proteomics projects. In this review we discuss critical issues related to data processing and analysis in proteomics and describe available methods and tools. We place special emphasis on the elaboration of results that are supported by sound statistical arguments.

Introduction

A main goal of proteomics has been the complete and in most cases quantitative analysis of the proteome of a species or, in multicellular organisms, a particular cell or tissue type. Although this goal has remained elusive, significant progress has been made in the development of an array of technologies for proteome analysis and their application to biological and clinical research[1]. At present, the vast majority of proteomic data are being generated by mass spectrometry, more specifically by tandem mass spectrometers of ever increasing performance[2]. These instruments and the diverse workflows they support have in common that they generate hundreds to tens of thousands of fragment ion spectra per hour of data acquisition. The assignment of these fragment ion spectra to peptide sequences, the inference of the proteins represented by the identified peptides and the determination of their abundances in the analyzed sample present complex computational and statistical challenges. It is essential for proteomics to develop and generally apply tools and solutions to these problems that provide accurate and reproducible results. Failure to do so introduces and propagates errors in the literature, makes it difficult for reviewers and readers to evaluate the conclusions of manuscripts or to meaningfully compare the results of different studies, and renders databases containing proteomic data essentially useless[3]. In this review we discuss critical problems facing the analysis of mass spectrometry–derived proteomic datasets and present the currently available solutions.

Assignment of fragment ion spectra to peptide sequences

The currency of information for tandem mass spectrometry (MS/MS) based proteomics is the fragment ion spectrum (MS/MS spectrum) of a specific peptide ion that is fragmented, typically in the collision cell of a tandem mass spectrometer. The correct assignment of such a spectrum to a peptide sequence is a first and central step in proteomic data processing. A large number of computational approaches and software tools have been developed to automatically assign peptide sequences to fragment ion spectra. These can be classified into three categories: (i) Database searching, where peptide sequences are identified by correlating acquired fragment ion spectra with theoretical spectra predicted for each peptide contained in a protein sequences database, or by correlating acquired fragment ion spectra with libraries of experimental MS/MS spectra identified in previous experiments (spectral library searching); (ii) *De novo* sequencing, where peptide sequences are explicitly read out directly from fragment ion spectra; and (iii) hybrid approaches, such as those based on the extraction of short sequence tags of 3–5 residues in length, followed by 'error-tolerant' database searching. For large-scale proteomics studies database searching remains the most frequently used peptide identification method. However, the other strategies provide attractive alternatives in specific situations, as discussed below.

[1]University of Michigan, Department of Pathology and Center for Computational Medicine and Biology, Ann Arbor, Michigan 48105, USA. [2]Purdue University, Departments of Statistics and Computer Science, West Lafayette, Indiana 47107, USA. [3]Institute of Molecular Systems Biology, Swiss Federal Institute of Technology (ETH) Zurich, CH-8093 Zurich, Switzerland and Faculty of Sciences, University of Zurich, CH-8006 Zurich, Switzerland. [4]Institute for Systems Biology, Seattle, Washington 98103, USA. Correspondence should be addressed to R.A. (rudolf.aebersold@imsb.biol.ethz.ch).

PUBLISHED ONLINE 27 SEPTEMBER 2007; DOI:10.1038/NMETH1088

Spectral identification by sequence database searching. Several MS/MS database search programs have been developed (**Table 1**), and their basic functionality is illustrated in **Figure 1**. The programs take the fragment ion spectrum of a peptide as input and score it against theoretical fragmentation patterns constructed for peptides from the searched database. The pool of candidate peptides is restricted based on user-specified criteria such as mass tolerance, proteolytic enzyme constraint and types of post-translational modification allowed (see **Supplementary Notes** online for discussion of the most important criteria). The output from the program is a list of fragment ion spectra matched to peptide sequences, ranked according to the search score. Typically, only the best scoring peptide match is considered during the subsequent statistical analysis step (see below). The search score measures the degree of similarity between the experimental spectrum and the theoretical spectrum, and therefore serves as the primary discriminating parameter for separating correct from incorrect identifications.

Table 1 | A list of publicly available tools for MS/MS-based proteomics

Program	Reference	Website
Database search tools		
SEQUEST	84	http://www.thermo.com
MASCOT	85	http://matrixscience.com[a]
ProteinProspector	86	http://prospector.ucsf.edu[a]
ProbID	87	http://tools.proteomecenter.org/wiki/index.php?title=Software:ProbID[b]
TANDEM	88	http://www.thegpm.org[a,b]
SpectrumMill		http://www.chem.agilent.com
Phenyx	89	http://www.phenyx-ms.com
OMSSA	4	http://pubchem.ncbi.nlm.nih.gov/omssa[a,b]
VEMS	90	http://personal.cicbiogune.es/rmatthiesen[b]
MyriMatch	91	http://www.mc.vanderbilt.edu/msrc/bioinformatics[b]
Spectral matching tools		
SpectraST	12	http://www.peptideatlas.org/spectrast
X! P3	92	http://p3.thegpm.org/tandem/ppp.html
Biblispec	11	http://proteome.gs.washington.edu/bibliospec
De novo sequencing tools		
Lutefisk	93	http://www.hairyfatguy.com/lutefisk[b]
PepNovo	94	http://peptide.ucsd.edu/pepnovo.py[a,b]
PEAKS	95	http://www.bioinformaticssolutions.com
Sequit		http://www.proteomefactory.com
Sequence tag/hybrid approaches		
GutenTag	16	http://fields.scripps.edu/GutenTag
Inspect	17	http://peptide.ucsd.edu/inspect.html[a,b]
Popitam	96	http://www.expasy.org/tools/popitam
Statistical validation of peptide and protein identifications		
PeptideProphet	21	http://www.proteomecenter.org/software.php[b]
ProteinProphet	56	http://www.proteomecenter.org/software.php[b]
Scaffold		http://www.proteomesoftware.com
Databases for storing and mining of mass spectrometry data		
PeptideAtlas	97	http://www.peptideatlas.org
Proteios		http://www.proteios.org
SBEAMS		http://sbeams.org
CPAS	98	https://www.labkey.org
PRIDE	99	http://www.ebi.ac.uk/pride
Data sharing		
Tranche		http://www.proteomecommons.org/dev/dfs[b]
Tools for protein quantification		
PEPPER (label free)	74	http://www.broad.mit.edu/cancer/software/genepattern[b]
EXPRES (isotopic labeling)		http://www.proteomecenter.org/software.php[b]
Libra (iTRAQ isobaric labeling)		http://www.proteomecenter.org/software.php[b]
ASAPRatio (label free)	100	http://www.proteomecenter.org/software.php[b]
MSQuant (isotopic labeling)		http://msquant.sourceforge.net[b]
RelEx (isotopic labeling)	101	http://fields.scripps.edu/relex

[a]Free access through the web interface (functionality might be limited). [b]Free distribution.

A number of scoring schemes have been described in the literature, including spectral correlation functions (for example, SEQUEST) or related concepts such as shared fragment counts and dot product (for example, TANDEM, OMSSA, MASCOT). Scoring functions can also be based on empirically observed rules (for example, SpectrumMill) or statistically derived fragmentation frequencies (for example, PHENYX). The score that is actually reported by the tool can be based on a somewhat arbitrary scale (for example, Xcorr score in SEQUEST), or converted to a statistical measure called expectation value, *E* value, which refers to the expected number of peptides with scores equal to or better than observed score under the assumption that peptides are matching the experimental spectrum by random chance (OMSSA, TANDEM and more recently MASCOT). *E* value is computed either by assuming that the database search score follows a certain (for example, Poisson) distribution[4,5], or by empirical fitting of the observed distribution of scores[6] (see **Fig. 1**). This score is largely invariant under different scoring methods and gives a clearer interpretation of goodness of match across different instrument platforms and search algorithms. It should be stressed, however, that neither the best match nor a high search score (or low *E* value) are reliable indicators for a true match. Discriminating true from false matches is therefore a critical next step in proteomic data analysis.

Spectral identification by spectral matching. A notable inefficiency of shotgun proteomics experiments lies in the repeated rediscovery of the same identifiable peptides by sequence database searching methods, which often are time consuming and error prone. With the availability of large amounts of proteomic data, part of which are organized in generally accessible databases (**Table 1**), it can be anticipated that all the proteins of a species that are detectable by mass spectrometry will eventually have been discovered. In fact,

systematic sequencing of proteins produced by microbes and eukaryotic species[7,8] has already reached remarkable depth of proteome coverage. Such extensive proteome maps now open the possibility of inferring the sequence of a peptide by matching its fragment ion patterns against a library of spectra representing the peptide sequences contained in the proteome map[9–12].

In this approach, a spectral library is compiled meticulously from a large collection of experimentally observed mass spectra of correctly identified peptides. An unknown spectrum can then be identified by comparing it to all the candidates in the spectral library to determine the match with the highest spectral similarity[13]. Recently, a number of tools have been developed that support peptide identification by spectral matching (**Table 1**). The spectral matching approach substantially outperforms classical sequence database searching in speed, error rate and sensitivity characteristics of the results[12] and has the advantage that the statistical models developed for assessing the output of database search tools (see below) are easily adaptable to the method[12]. However, no peptides will be identified that were not previously entered into the respective spectral library. At this time, when no proteome map has been completed, spectral matching approaches might be used most effectively as a rapid first pass in an incremental search strategy.

Spectral identification by *de novo* sequencing. In the *de novo* sequencing approach the amino acid sequence of a peptide is explicitly read from a fragment ion spectrum. Initially this was accomplished manually. More recently, an array of tools has been developed that assist the researcher with this task (**Table 1**). The main advantage of *de novo* sequencing over the database search method is that it allows identification of spectra for which the exact peptide sequence is not present in the searched sequence database, such as peptides containing sequence polymorphisms and modified peptides. It is therefore mainly used for protein analysis in species for which no or limited genome sequence information is available or for identifying modified peptides. However, *de novo* analysis is computationally intensive and requires high quality fragment ion spectra. Furthermore, researchers analyzing proteomic data are more interested in knowing what proteins are present in the sample. This means that peptide sequences extracted from MS/MS spectra using *de novo* algorithms need to be matched, using for example BLAST, against the sequences of known proteins present in the sequence databases, a strategy that is tedious in high throughput proteomics environment. Thus, a more effective strategy may be to start with database searching, and apply *de novo* sequencing tools to the remaining unassigned high quality spectra[14].

Spectral identification with hybrid approaches. Spectral identification can also be carried out using hybrid approaches that combine elements of both *de novo* sequencing and database searching. The analysis starts with inference of short sequence tags (partial sequences) from MS/MS spectra, followed by an error-tolerant database search: that is, a search that allows one or more mismatches between the sequence of the peptide that produced the MS/MS spectrum and the database sequence. First pioneered in ref. 15, this approach has been recently extended by several groups[16,17] (see **Table 1**). By limiting the search space to only those database peptides that contain the sequence tag extracted from the spectrum (or one of the several sequence tags, if more than one per spectrum is extracted), the database search time can be significantly reduced. Hybrid approaches are also potentially very powerful for the systematic analysis of post-translationally modified peptides, or peptides containing artifactual modifications. Allowing all possible types of modifications at all possible sites leads to a combinatorial explosion of the database search space and is therefore poorly compatible with sequence database searching. The use of sequence tags, or related approaches such as look-up peaks[18] can reduce the size of the space to be searched back to manageable levels.

Statistical assessment of peptide assignments in large-scale datasets

Database and spectral matching search tools typically produce a peptide match for each input spectrum, some of which may be true matches and some false. In some experiments, the best-scoring peptide assignment produced by a database search program is incorrect for the majority of searched MS/MS spectra. Some of the reasons for the high failure rate are listed in **Supplementary Notes**. Early on in proteomics it was customary to generate a list of 'high confidence' identifications according to an *ad hoc* cutoff value of the score provided by the search engine, often in conjunction with visual inspection of peptide assignments to fragment ion spectra by an expert. However, the score distributions produced by a search tool depend on a multitude of factors, including the performance of the mass spectrometer, data quality, and the size of the database. Thus, application of the same thresholds to data from different experiments would result in different (and unknown) error rates, making comparison between datasets practically impossible. Manual inspection by an expert cannot be regarded as viable validation process because it is time-consuming and not compatible with the high numbers of fragment ion spectra acquired in proteomics, it is subjective, and the results depend on the level of expertise of the validating individual. Therefore, modern

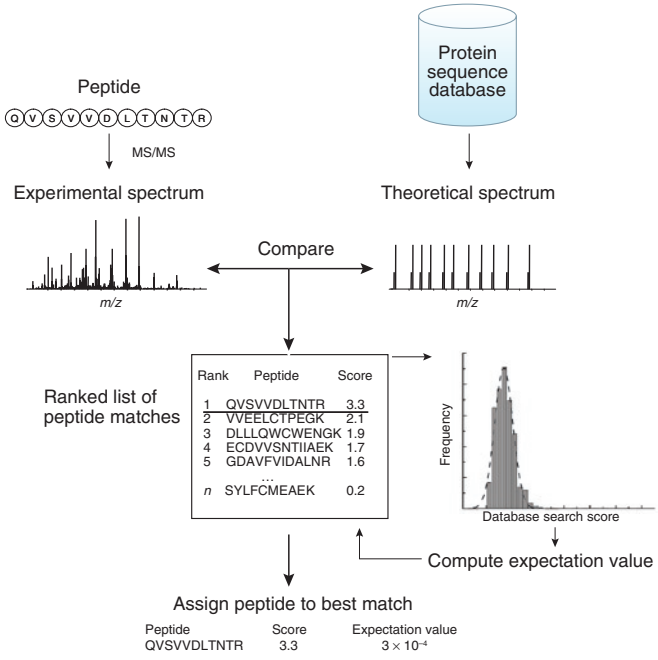

Figure 1 | Peptide identification by MS/MS database searching. An acquired MS/MS spectrum is correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the spectra. Candidate peptides are ranked according to the computed search score, and the highest scoring peptide sequence is selected for further analysis.

proteomics has gradually moved away from manual inspection of the data and *ad hoc* scoring schemes, and toward probabilistic approaches that provide statistical measure of confidences and estimates of error rates. See **Box 1** and **Table 2** for statistical terminology relevant to the assessment of database search results.

Recently, several approaches that translate the database search tool output scores into probabilities or estimated false discovery rates (FDRs)[19] have been introduced. These global approaches (as illustrated in **Fig. 2**) are concerned with modeling the distribution of search scores constructed by taking the top-scoring peptide assignment for each experimental spectrum in the whole dataset ('global distribution'). This distinguishes them from the expectation value calculation involving modeling the single-spectrum distribution of scores constructed for each experimental spectrum separately from all peptides in the searched sequence database that were scored against that particular spectrum. In fact, the global and single-spectrum–based approaches are complementary: that is, whole-dataset modeling and FDR analysis can be performed using *E* values in place of the original search scores. The global statistical approaches can be broadly grouped into two categories: target-decoy searching and empirical Bayes approaches.

Target-decoy searching. The methods of the first group rely solely on searching target-decoy databases, and compute an optimized cut-off score for each dataset. The target-decoy search strategy[20] involves two steps. In the first step MS/MS spectra are searched against a target database of protein sequences augmented with the reversed (or randomized, or shuffled) sequences of the same database. The approach assumes that matches to decoy peptide sequences and false matches to sequences from the original database follow the same distribution. The plausibility of these assumptions is discussed in ref. 20. In the second step, peptide assignments are filtered using various score cut-offs, and the corresponding FDR for each cut-off is estimated as $2N_d/N$, where N is the number of peptide matches with scores above the cut-off and N_d is the number of matches to decoy sequences among them.

The advantage of this FDR estimation method is that it is simple to implement and requires minimal distributional assumptions, which makes it easily applicable in a variety of situations. The drawbacks of this approach include doubling the database search time. A more fundamental issue arises of whether reversing or randomizing sequences can provide an accurate assessment of the distribution of false peptide matches when many of those are known to be sequences homologous to the true peptides rather than completely random sequences.

Empirical Bayes approaches. The methods in the second category are exemplified by PeptideProphet[21], which employs a so-called empirical Bayes[22] approach that models the distributions of database search scores and auxiliary information (see below) observed for all peptide assignments in the dataset as a two-component mixture of distributions representing correct and incorrect identifications. Before that step, PeptideProphet combines multiple search score–related parameters (for example, the search score itself, Xcorr, and its derivative, ΔC_n score, in the case of SEQUEST) into a single score, called discriminant search score. The discriminant score coefficients and the functional form of the resulting discriminant score distributions are determined for each search engine using training datasets. Those distributions are modeled, however, anew for each dataset using the expectation-maximization algorithm, leading to posterior probabilities of correct identifications as inferential indicators. These probabilities are then used to estimate the FDR for any minimum probability used as a cut-off. In contrast to the target-decoy database search approach, appending decoys is not necessary for deriving the distribution of incorrect identifications. Furthermore, additional modeling in PeptideProphet results in an increase of statistical power compared to threshold-based approaches.

The limitations of PeptideProphet are largely related to the parametric assumptions and, to a lesser degree, to the use of fixed

BOX 1 TERMINOLOGY AND GENERAL STATISTICAL METHODS FOR CONTROLLING FALSE DISCOVERY RATE

One can view spectral identification as a process of hypothesis testing in which the hypothesis H_0 'random chance identification' is tested for each spectrum against the alternative hypothesis H_a 'correct identification'.

Table 2 summarizes the outcome of identification of m MS/MS spectra. Counts U, V, T and S are unknown and random due to the stochastic nature of mass spectra. The total number of incorrectly identified spectra m_0 is unknown but fixed. Although V, the number of false positive identifications, is unknown, it is possible to estimate or bound various error rates that involve the expected value of V (that is, the average value that one would obtain after an infinite repetition of the experiment):

- False positive rate (FPR), or type I error, is a property of a single MS/MS spectrum, and is defined as the probability that a randomly matched spectrum is judged correct: FPR = $E(V)/m_0$.
- Family-wise error rate (FWER) is a property of m MS/MS spectra, and is defined as the probability of making at least one incorrect identification among all identifications judged correct: that is, FWER = $p(V \geq 1)$. Example of method controlling FWER: Bonferroni[102].
- False discovery rate (FDR) is a property of m MS/MS spectra, and is defined as the expected proportion of incorrect identification among all identifications judged correct: that is, FDR = $E(V)/R$. Examples of methods controlling FDR: step-up[19], permutation-based[103], empirical Bayes[22,104].

Table 2 | Outcomes of applying a classification rule

	No. of matches judged incorrect	No. of matches judged correct	Total
Number of truly incorrect matches	U	V	m_0
Number of truly correct matches	T	S	$m - m_0$
Total	$m - R$	R	m

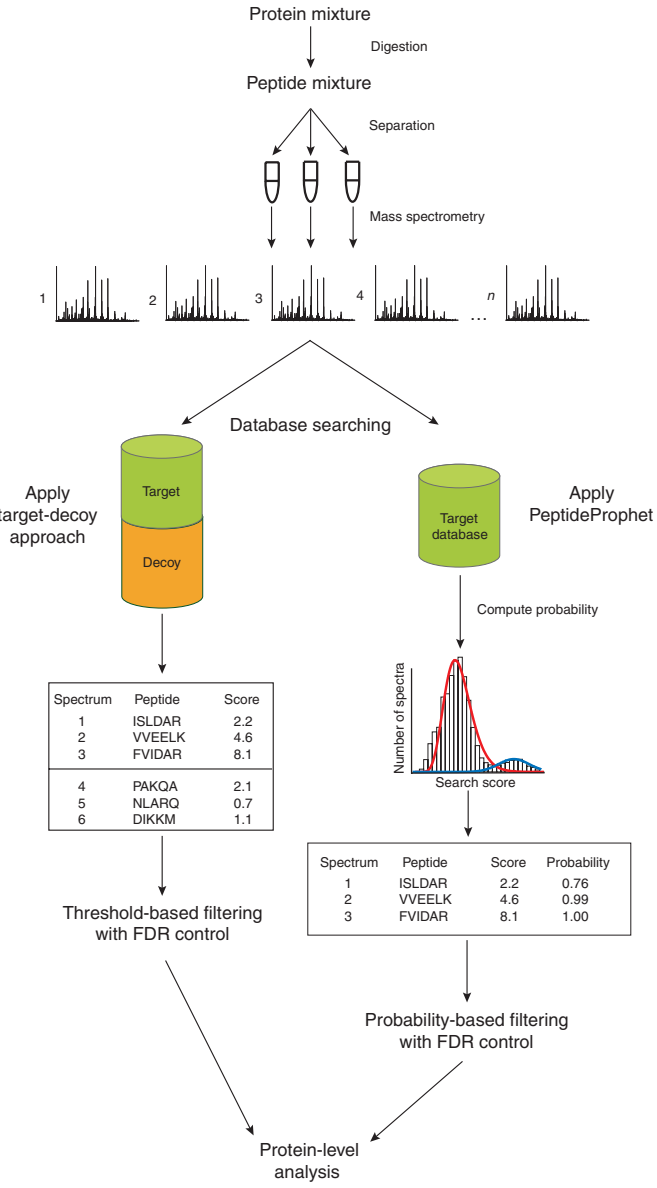

Figure 2 | Statistical analysis of large-scale datasets of peptide assignments. In the target-decoy strategy (left), all spectra from the entire experiment are searched against a composite target plus decoy database, and then the numbers of matches to decoy peptides are used to estimate the false discovery rate (FDR) resulting from filtering the data using various score thresholds. In the probabilistic mixture-modeling approach (right), the most likely distributions among correct (red curve) and incorrect (blue curve) peptide assignments are fitted to the observed data (histogram). A probability is computed for each peptide assignment in the dataset, which can then be used to estimate the FDR.

coefficients in computing the discriminant search score. These limitations are currently being addressed in several ways. First, the probabilistic modeling approach of PeptideProphet and the decoy strategy can be combined within a single framework through a semi-supervised expectation-maximization algorithm that explicitly incorporates the class label available for decoy peptide matches. Second, parametric specification of continuous mixture components in the model can be relaxed, for example, by using multiple components to model each class of peptides, correct and incorrect. These new developments result in improved robustness and higher accuracy of computed probabilities even in the case of the most challenging datasets (H. Choi and A.I.N., unpublished data).

Which scoring method when? The statistical power of all the identification procedures is strongly influenced by a number of factors, including the discriminative ability of the database search score, the quality of the spectra and the size of the database. Although there is currently no theory on the optimality of the score, empirical evidence suggests that some scores perform better than others in different settings[23,24]. Combining several search scores produced by the same search tool improves the overall performance[21,25–28]. Several programs (for example, TANDEM and SpectrumMill) allow an efficient multi-step analysis, starting with an enzyme-constrained search, followed by a second search for peptides with modifications, nonspecific cleavage or missed cleavage sites. The statistical power of the identification procedure can also indirectly benefit from further processing of MS/MS spectra performed before database search[29,30], clustering of redundant spectra[31,32], recognition of spectra produced by cofragmentation of two or more peptides[33], removal of low quality spectra[14,34–37] and application of automated charge state determination algorithms[38,39]. Furthermore, improved discrimination can be achieved by combining the output from two or more different database search tools[40–43] or by combining data from multiple consecutive stages of mass spectrometry (for example, MS/MS and MS/MS/MS (MS^3))[44].

Use of auxiliary information to improve spectral identification

The database search score (or the composite of multiple scores) measuring the degree of similarity between the experimental and theoretical spectra represents only one set of discriminant features useful for separating correct from incorrect identifications. Using this information alone, it may be difficult to accurately separate true from false identifications, even if optimal statistical methods are being used. The discrimination can be further improved if auxiliary information that may be generated coincidentally in the course of a proteomics experiment is also included in the analysis. Such types of information include mass accuracy—that is, the difference between the measured and calculated mass of the peptide ions (available from the first stage of mass spectrometry, MS^1)—and peptide separation coordinates, for example, retention time[45,46] or pI value[47–49] (peptide separation step). Other useful peptide properties include the number of termini consistent with the type of enzymatic cleavage used and the number of missed cleavage sites (digestion step). In some cases, additional information such as presence of a specific amino acid or sequence motif—for example, cysteine in the case of avidin affinity purification of peptides containing biotinylated cysteines[1], or the sequence motif N-X-S/T for peptides containing *N*-linked glycosylation sites[50] (peptide enrichment step)—can be used as further constraints.

These types of auxiliary information provide evidence that can be used to incrementally augment the search score(s) generated by the search engine. The availability and information content depends strongly on the experiment that was carried out to generate the data. For example, the contribution of the mass accuracy parameter to differentiating true from false identifications depends on the mass accuracy of the mass spectrometer used, and the pI value is only useful if isoelectric focusing was used as one of the peptide separation tools. Although it is possible to take into account auxiliary information in the threshold-based approaches[46,49,51–53], handling experimental

variations (for example, a bias in the mass measurement, or inaccurate determination of the pH value in each peptide fraction) can be problematic. Application of threshold-based approaches also requires datasets of sufficiently large size owing to the need to subdivide peptide assignments into subcategories based on the search score and all extra parameters. At the same time, auxiliary information can be effectively used in PeptideProphet[21,47,54]. As it models all data types simultaneously, it has the inherent flexibility to detect and correct for measurement bias, and to weigh the contributions of the different types of information as a function of the experiment in computing posterior peptide probabilities.

Inferring protein identifications from spectral identifications

The purpose of most proteomic experiments is not the identification of peptides, but the identification of the proteins present in the sample before digestion[55]. Thus, the peptide sequences of the identified fragment ion spectra need to be grouped according to their corresponding protein, and the confidence measures need to be recomputed at the level of proteins. This process is not straightforward owing to several challenges, and it is a likely source of significant errors in the proteomics literature.

The first challenge is related to the fact that many correctly identified peptides tend to group into a relatively small number of proteins[56]. This is particularly obvious in the analysis of human serum samples, where the dominant peptide identifications come from a dozen of the most abundant serum proteins, and the total number of identified proteins is typically less than a thousand[57]. At the same time, incorrect spectral identifications match randomly to the much larger number of proteins in the searched sequence database (for example, more than 40,000 in human IPI database; **Box 2**). Thus, almost every high-scoring incorrect spectral assignment introduces one additional incorrect protein identification, resulting in an increase in the false discovery rates when going from the spectral to the protein level.

The second challenge arises because of shared peptides: that is, peptides whose sequence is present in more than a single entry in the protein sequence database. Such cases most often result from the presence of homologous proteins, splicing variants or redundant entries in the protein sequence database. This problem is particularly serious in the case of higher eukaryote organisms[55,58]. As a result, in shotgun proteomics it is often not possible to differentiate between different protein isoforms. In general, this is less of a problem when proteins are first separated using a multidimensional protein separation technique (for example, using two-dimensional gels), where additional information such as the molecular weight of the sample proteins can assist in the determination of the protein identities. A detailed discussion of the difficulties in interpreting the results of shotgun proteomics experiments at the protein level can be found in ref. 55.

Most frequently, protein identification is performed by determining peptide sequence identity in MS/MS spectra as described in the previous section, and by grouping peptide sequences into proteins, deterministically[40,59,60] or probabilistically (for example, by apportioning peptides to proteins with some weights[41,55,56]). An alternative approach[61] sidesteps the process of spectral identification, and combines overlapping uninterpreted MS/MS spectra into longer chains, then maps the chains to protein sequences directly. With both approaches, combining MS/MS spectra into proteins is often insufficient for unambiguous protein identification owing to a large number of shared peptides, in particular in cases when the protein database contains many homologous proteins and isoforms. Thus the issue is what it means for a protein to be identified. Some publications report all proteins identified with at least one distinct peptide, or select one representative protein among isoforms and homologs[62]. A nomenclature based on the parsimony principle (also called Occam's razor), which consists of determining the smallest number of proteins that can account for all observed peptides, has been described in ref. 55 and provides a consistent and concise way of representing the results of a proteomic experiments.

Once peptides are grouped into proteins, the plausibility of the protein identification is quantified with a score. On one hand, protein identifications with low spectral coverage are likely to be spurious. On the other hand, the number of identified peptides mapped to a protein sequence is strongly correlated with length and abundance of the

BOX 2 SEQUENCE DATABASES

The most commonly used protein sequence databases for searching MS/MS spectra include

1. Entrez Protein database from the US National Center for Biotechnology Information (NCBI)
2. Reference Sequence (RefSeq) database from NCBI
3. UniProt, consisting of Swiss-Prot and its supplement, TrEMBL
4. International Protein Index (IPI) database, maintained by the European Bioinformatics Institute

Databases vary in terms of their completeness, degree of redundancy and quality of sequence annotation. Entrez Protein is the most complete database; however, it contains many redundant sequences (partial mRNAs, sequencing errors and so forth), and the entries are not as accurately annotated as those in Swiss-Prot or RefSeq. For the six organisms for which it is available, the IPI database represents a good balance between completeness and degree of redundancy. It also maintains cross-references to all its source data (Ensembl, UniProt, RefSeq), making biological data interpretation easier.

Genomic databases can also be used for MS/MS database searching[105,106]. This is an attractive option for the identification of peptides not yet present in any protein sequence database: for example, previously unidentified alternative splice forms, or sequence polymorphisms. The search can be conducted against translated expressed sequence tag (EST) databases, or against the DNA sequence translated in all six frames. Alternatively, a database of putative splice forms can be created using computational gene prediction models. Searching genomic databases should be practiced with great caution, as accurate translation to protein sequence is complicated owing to frame-shifts, incorrectly predicted open reading frames, sequencing errors and so forth. Such searches are also computer intensive, although several recent studies describe efficient computational solutions[107]. Thus, such searches should be done using only high quality MS/MS spectra that could not be identified by a normal search against a protein sequence database[14].

protein, and one can hardly expect good peptide coverage in complex mixtures, or with experimental designs that enrich for a particular class of peptides. Scoring functions attempt to distinguish between false and true protein identifications in a number of ways. For example, the Bayes rule–based scoring scheme in ProteinProphet includes the concept of the number of sibling peptides[56]. Other approaches are based on Poisson distribution–based statistics that take into account the protein length[62,63] or model the protein abundance as a latent variable[41].

The final goal of the protein-level analysis is to derive a list of proteins with a controlled FDR. FDR-controlling procedures that are analogous to the ones used at the spectral level are frequently used for proteins. For example, protein-level FDR can be again estimated using the target-decoy strategy[20,60], or as a sum of posterior probabilities of correct identifications[56]. In addition, *P* values can be derived directly from the distributional assumptions of protein identification followed by Bonferroni adjustment to control family-wise error rate[62], a more conservative criterion than FDR.

As in the case of spectral-level analysis, the statistical power of protein identification depends on the scoring function and the method used to control FDR. The power can be improved by incorporating more information into the scoring function: for example, predicted detectability of peptides[64] or similarity of quantitative profiles of peptides mapped to a same protein. As the level of analysis (MS/MS spectra, distinct peptide sequences, proteins, and so forth) and the methods used to compute data summaries at each level become more complex, proving the appropriateness of data analysis procedures becomes more difficult. At the protein level, the ultimate validation of the results can be obtained by independent technical and biological replication of the experiment using the same or a different (for example, targeted) experimental strategy.

Quantitative proteomics

Mass spectrometry is increasingly used for relative or absolute quantification of peptides and proteins[1,65]. A typical analysis involves extraction of quantitative information from mass spectra at various levels of summarization, such as MS^1 spectrum features (peaks in the MS^1 spectrum characterized by their intensity, *m*/*z* value, and the time of acquisition of the spectrum), peptide features (that is, groups of isotopic mass peaks originating from the same peptide ion), or peptide (that is, multiple peptide features corresponding to different charge states of the same peptide). The goal of the experiment is to quantify changes in the abundance of those features across the samples that are being compared, and to provide a maximal list of differentially abundant features with a controlled FDR. Quantitative proteomics workflows can be generally divided into three categories (**Fig. 3**): stable isotope labeling, spectral counting and spectral feature analysis.

Stable isotope labeling. One commonly used approach is based on stable isotope labeling of proteins, in which samples are labeled chemically (for example, in isotope coded affinity tag, ICAT; or isobaric tags for relative and absolute quantification, iTRAQ) or metabolically (for example, in stable isotope labeling with amino acids in cell culture, SILAC), mixed together, and digested into peptides[1,65] (**Fig. 3a**). Because of the mass shift introduced by the reagent, MS^1 spectrum features corresponding to the same peptide can be quantified separately in the same mass spectrometry run, and their ratio represents the relative abundance of the corresponding peptide. Representative tools supporting this type of analysis are listed in **Table 1**. The correspondence between the spectral features representing the same peptide is established through the identification of the peptide sequence from acquired MS/MS spectra. In addition to the increased complexity posed by the labeling steps, this workflow is limited by the need to acquire and interpret the MS/MS spectra.

Spectral counting. Quantification can also be done without isotopic labeling by means of spectrum counting (from MS/MS data) or integrated ion intensities (MS^1)[62,66–70] (**Fig. 3b**). In this strategy, the samples that are being compared are analyzed in the mass spectrometer separately but using the same data acquisition protocol. A separate list of proteins is created for each of the samples, and the lists are then compared to find differentially expressed proteins. The protein abundance in each sample is estimated from the number of MS/MS spectra identified corresponding to each protein normalized to account for protein length or expected number of tryptic peptides. As a variation of this strategy, peptide abundance can be determined from the intensity of the corresponding spectrum features. This method suffers from inability to quantify low abundance proteins identified from only one or two peptides, and in general is less accurate than the methods based on stable isotope labeling. Still, the practical utility of this method has been demonstrated in a number of applications[66–69,71,72].

Spectral feature analysis. The third kind of the workflow is different from the first two in that it does not require identification of the peptide sequence corresponding to each observed spectrum feature before quantification[50,73] (**Fig. 3c**). In this label-free strategy, biological samples are analyzed in separate mass spectrometry runs, and the correspondence between spectral features across the runs is established by means of computational tools and with at most a minimal amount of information from MS/MS spectra[50,73–75]. This workflow allows analysis of a large number of spectrum features and allows higher data throughput, and is compatible with applications that require profiling of multiple biological samples, such as proteomics-based candidate biomarker discovery. The drawbacks include increased computational complexity owing to the presence of a large number of spurious features and noise, and more stringent requirements in robustness and reproducibility at various data acquisition steps[76,77]. Subsequent or parallel experiments using, for example, targeted workflow[2,78] are typically necessary to verify the presence of these features and their changes in abundance, and to determine their identity.

Regardless of the workflow used, a typical output consists of a list of detected proteins (or spectral features for which the identity may not be known) and their absolute or relative abundances across all samples or runs. The resulting information is similar to the information from other high-throughput experiments, such as gene expression microarrays. Determining changes in abundance that are significant requires statistical methods that take advantage of the large number of features to compensate for the small sample size[76]. A number of such methods have been implemented as a part of the computational tools that perform quantification. Alternatively, data can be exported to an external tool developed more generally in support of high-throughput data processing; for example, those that are available as a part of the Bioconductor project[79]. Furthermore, the methods described in **Box 1** can be used to control the FDR in the list of differentially abundant features.

Although data from quantitative proteomic experiments have similarities to other data, such as those from gene expression experiments, they present many specific challenges. Proteomic data are more

complex than gene expression data owing to the large span of protein concentrations, and to the fact that the identities (peptide sequences) of spectral features are not always known, or may be determined incorrectly. Another complication relates to the ambiguity in assigning peptides to proteins[55]. In the case of a shared peptide, its quantification may not be a reliable measure of the abundance of any of its corresponding proteins. In fact, the procedures for peptide identification and quantification are interdependent and complementary, and the power of both procedures can be increased by summarizing the data at different levels, such as at the level of protein identity. For

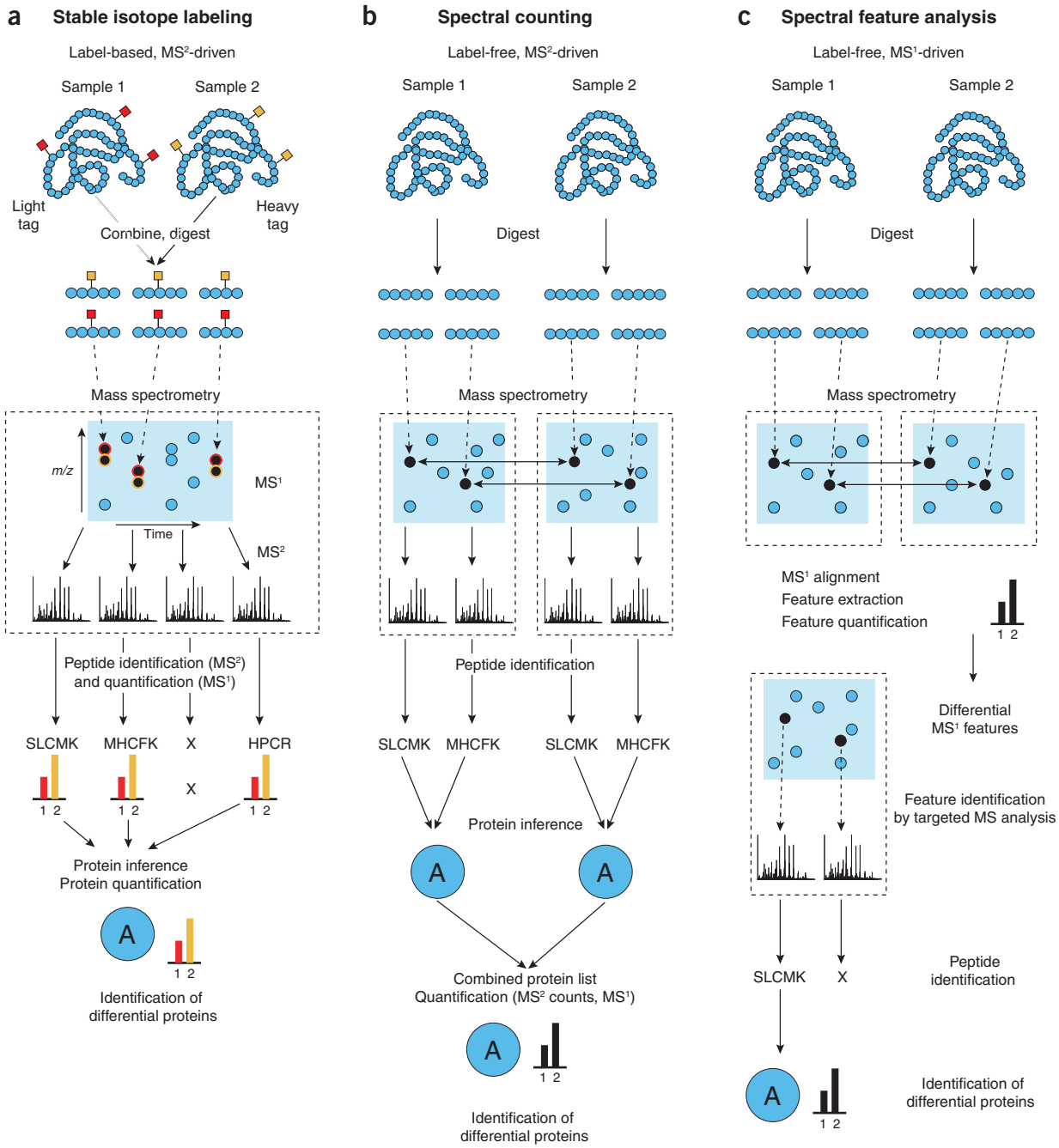

Figure 3 | Quantitative proteomics workflows. (**a**) In the stable isotope labeling workflow, proteins are labeled using a light (sample 1, red) or heavy (sample 2, yellow) mass tag, mixed, digested into peptides and analyzed using tandem mass spectrometry. Spectral features observed in MS^1 data (indicated as black dots in the $m/z \times$ retention time plots) are identified from the acquired MS/MS spectra. Identified peptides are quantified from the signal intensities of MS^1 features, and this information is used to infer the identity and relative quantification of their corresponding protein (protein A). Spectral features for which no MS/MS spectrum was acquired (blue dots), or for which no high probability peptide assignment was obtained (indicated by X) are not further analyzed. (**b**) In the spectral counting strategy, unlabeled protein samples are analyzed separately using the same protocol as each other, and the relative protein quantification is established by comparing the number of MS/MS spectra identified for each protein. (**c**) In the spectral feature analysis strategy, the analysis starts with alignment of MS^1 data from different samples, extraction of spectral features and their quantification, all of which is done before the identification step. Spectral features showing differential expression are then identified using a targeted MS/MS-based workflow.

BOX 3 MASS SPECTROMETRY DATA FORMATS

Reflecting the diversity of mass spectrometry instrumentation, experimental platforms and computational tools, a number of different file formats exist and are in active use by proteomics researches[108,109]. The 'raw' data are acquired by a mass spectrometer and stored in a proprietary binary file format such as Xcalibur/RAW (Thermo Electron), Analyst/WIFF (ABI and MDS Sciex), MassLynx/RAW (Waters) or BAF (Bruker). Before database searching, MS/MS spectra need to be extracted from raw data in plain text format: for example, mgf (MASCOT), dta (SEQUEST) or pkl (ProteinLynx, Micromass) files. The output from the search engines also varies from simple text files with a minimal amount of information (for example, out files in SEQUEST) to more comprehensive files (for example, XML-based files in TANDEM). More recently, several open file formats have been introduced that simplify exchange and analysis of proteomic data, such as mzXML and mzDATA for storing mass spectrometry data, and pepXML and protXML[43] for peptide and protein-level identification results. Software is now available for converting raw data to mzXML[108] and mzDATA, between these two formats, and from them to search engine–specific input file formats. A more detailed list of existing file formats and conversion utilities can be found at http://tools.proteomecenter.org/software.php and http://www.proteomecommons.org. A new format unifying mzXML and mzDATA, mzML, is expected to become available in early 2008. More ambitious efforts to develop comprehensive standards and data ontologies for proteomics are under way under the umbrella of the HUPO Proteomics Standards Initiative (http://www.psidev.info).

example, shared quantitative profiles of peptides corresponding to the same protein increase the confidence in the identification. Conversely, observing changes in abundance across different peptides from the same protein may suggest the presence of several protein isoforms having differential expression[55].

Conclusions and outlook

Mass spectrometry–based proteomics, specifically proteome analysis by a shotgun approach, has reached a high level of maturity with respect to sample processing, data acquisition and data analysis. However, a number of significant challenges remain. They are primarily related to the complexity of proteomes, which has so far precluded true proteomic analyses (that is, the analysis of all the components of a proteome) and generated partially overlapping datasets from identical samples, suggesting poor reproducibility of the technology. Secondarily, these challenges are related to the analysis of the information contained in proteomic datasets. In combination, these problems have created the impression that published proteomic data are at times of dubious quality.

It can be expected that incremental improvements of tools and methods such as the ones described in this review will further increase the quality of published proteomics data. The most significant improvements, however, will come from the skilled and systematic application of the most advanced available tools. It is encouraging that leading journals publishing proteomic studies have recognized this fact and started to request that authors follow specific guidelines and that the raw data supporting the conclusions of a paper be made accessible. The practical implementation of these guidelines is facilitated by the development of data sharing mechanisms such as Tranche (**Table 1**) and common file formats (**Box 3**). We must recognize, however, that some of the informatics issues facing shotgun proteomics datasets can be completely resolved by neither expert validation nor statistical arguments. These include the problem of inferring the identities of the proteins, protein isoforms and differentially modified proteins in a sample from confidently identified peptides. This and similar problems, in our opinion, can only be rigorously solved by the development of alternative proteomic workflows.

Two such alternatives are becoming apparent. The first, referred to as top-down proteomics, is focused on the analysis of intact proteins rather than peptides and therefore has the potential to resolve populations of proteins into their components[80–82]. The second alternative is based on targeted analysis of specific peptides of high information content, termed proteotypic peptides, that collectively represent the proteome, thus eliminating to a large extent the redundancy of current methods[2,64,83]. Although substantial progress has been achieved in both directions, significant technology development, including development of new algorithms and analysis tools, remains before the routine implementation of these technologies.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported in part by US National Institutes of Health (NIH) National Cancer Institute Grant R01 CA126239 to A.I.N. and with federal funds from the National Heart, Lung, and Blood Institute of the NIH under contract no. N01-HV-28179 to R.A.

Published online at http://www.nature.com/naturemethods
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212–217 (2006).
3. Carr, S. *et al.* The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3**, 531–533 (2004).
4. Geer, L.Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
5. Sadygov, R.G. & Yates, J.R. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75**, 3792–3798 (2003).
6. Fenyo, D. & Beavis, R.C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75**, 768–774 (2003).
7. King, N.L. *et al.* Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol. [online]* **7**, R106 (2006).
8. Brunner, E. et al. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* **25**, 576–583 (2007).
9. Yates, J.R., Morgan, S.F., Gatlin, C.L., Griffin, P.R. & Eng, J.K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* **70**, 3557–3565 (1998).
10. Craig, R., Cortens, J.C., Fenyo, D. & Beavis, R.C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **5**, 1843–1849 (2006).
11. Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S. & MacCoss, M.J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684 (2006).
12. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
13. Stein, S.E. & Scott, D.R. Optimization and testing of mass-spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
14. Nesvizhskii, A.I. *et al.* Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient

identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **5**, 652–670 (2006).
15. Mann, M. & Wilm, M. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
16. Tabb, D.L., Saraf, A. & Yates, J.R. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **75**, 6415–6421 (2003).
17. Tanner, S. *et al.* InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626–4639 (2005).
18. Bern, M., Cai, Y.H. & Goldberg, D. Lookup peaks: a hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **79**, 1393–1400 (2007).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
20. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
21. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
22. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
23. Kapp, E.A. *et al.* An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* **5**, 3475–3490 (2005).
24. Elias, J.E., Haas, W., Faherty, B.K. & Gygi, S.P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2**, 667–675 (2005).
25. Lopez-Ferrer, D. *et al.* Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal. Chem.* **76**, 6853–6860 (2004).
26. Anderson, D.C., Li, W.Q., Payan, D.G. & Noble, W.S. 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.* **2**, 137–146 (2003).
27. Kislinger, T. *et al.* PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2**, 96–106 (2003).
28. Ulintz, P.J., Zhu, J., Qin, Z.H.S. & Andrews, P.C. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol. Cell. Proteomics* **5**, 497–509 (2006).
29. Gentzel, M., Kocher, T., Ponnusamy, S. & Wilm, M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* **3**, 1597–1610 (2003).
30. Mujezinovic, N. *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* **6**, 5117–5131 (2006).
31. Beer, I., Barnea, E., Ziv, T. & Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4**, 950–960 (2004).
32. Tabb, D.L., Thompson, M.R., Khalsa-Moyers, G., VerBerkmoes, N.C. & McDonald, W.H. MS2Grouper: Group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J. Am. Soc. Mass Spectrom.* **16**, 1250–1261 (2005).
33. Zhang, N. *et al.* ProblDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **5**, 4096–4106 (2005).
34. Moore, R.E., Young, M.K. & Lee, T.D. Method for screening peptide fragment ion mass spectra prior to database searching. *J. Am. Soc. Mass Spectrom.* **11**, 422–426 (2000).
35. Wong, J.W.H., Sullivan, M.J., Cartwright, H.M. & Cagney, G. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics [online]* **8**, 51 (2007).
36. Flikka, K., Martens, L., Vandekerckhoe, J., Gevaert, K. & Eidhammer, I. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* **6**, 2086–2094 (2006).
37. Xu, M. *et al.* Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. *J. Proteome Res.* **4**, 300–305 (2005).
38. Colinge, J., Magnin, J., Dessingy, T., Giron, M. & Masselot, A. Improved peptide charge state assignment. *Proteomics* **3**, 1434–1440 (2003).
39. Tabb, D.L. *et al.* Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra. *J. Am. Soc. Mass Spectrom.* **17**, 903–915 (2006).
40. Resing, K.A. *et al.* Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76**, 3556–3568 (2004).
41. Price, T.S. *et al.* EBP, a program for protein identification using multiple tandem mass spectrometry data sets. *Mol. Cell. Proteomics* **6**, 527–536 (2007).
42. Higgs, R.E. *et al.* Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. *J. Proteome Res.* **6**, 1758–1767 (2007).
43. Keller, A., Eng, J., Zhang, N., Li, X.-J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol. [online]* **1**, E1–E8 (2005).
44. Olsen, J.V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA* **101**, 13417–13422 (2004).
45. Strittmatter, E.F. *et al.* Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* **3**, 760–769 (2004).
46. Qian, W.J. *et al.* Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **4**, 53–62 (2005).
47. Malmstrom, J. *et al.* Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **5**, 2241–2249 (2006).
48. Xie, H. & Griffin, T.J. Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics. *J. Proteome Res.* **5**, 1003–1009 (2006).
49. Cargile, B.J., Bundy, J.L., Freeman, T.W. & Stephenson, J.L. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* **3**, 112–119 (2004).
50. Zhang, H. *et al.* High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. *Mol. Cell. Proteomics* **4**, 144–155 (2005).
51. Heller, M. *et al.* Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4**, 2273–2282 (2005).
52. Olsen, J.V. *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021 (2005).
53. Rudnick, P.A., Wang, Y.J., Evans, E., Lee, C.S. & Balgley, B.M. Large scale analysis of MASCOT results using a mass accuracy-based THreshold (MATH) effectively improves data interpretation. *J. Proteome Res.* **4**, 1353–1360 (2005).
54. Nesvizhskii, A.I. & Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics data sets generated by tandem MS. *Drug Discov. Today* **9**, 173–181 (2004).
55. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
56. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
57. Omenn, G.S. *et al.* Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core data set of 3020 proteins and a publicly-available database. *Proteomics* **5**, 3226–3245 (2005).
58. Rappsilber, J. & Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **27**, 74–78 (2002).
59. Yang, X. *et al.* DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* **3**, 1002–1008 (2004).
60. Weatherly, D.B. *et al.* A heuristic method for assigning a false-discovery rate for protein identifications from mascot database search results. *Mol. Cell. Proteomics* **4**, 762–772 (2005).
61. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P.A. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. USA* **104**, 6140–6145 (2007).
62. States, D.J. *et al.* Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **24**, 333–338 (2006).
63. Sadygov, R.G., Liu, H.B. & Yates, J.R. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **76**, 1664–1671 (2004).
64. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
65. Goshe, M.B. & Smith, R.D. Stable isotope-coded proteomic mass spectrometry. *Curr. Opin. Biotechnol.* **14**, 101–109 (2003).
66. Old, W.M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4**, 1487–1502 (2005).
67. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**, 1265–1272 (2005).

68. Zybailov, B., Coleman, M.K., Florens, L. & Washburn, M.P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* **77**, 6218–6224 (2005).
69. Liu, H., Sadygov, R.G. & Yates, J.R. III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
70. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P.C. & Geromanos, S.J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
71. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).
72. Blondeau, F. *et al.* Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc. Natl. Acad. Sci. USA* **101**, 3833–3838 (2004).
73. Radulovic, D. *et al.* Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3**, 984–997 (2004).
74. Jaffe, J.D. *et al.* PEPPeR, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics* **5**, 1927–1941 (2006).
75. Li, X.-J., Yi, E.C., Kemp, C.J., Zhang, H. & Aebersold, R. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* **4**, 1328–1340 (2005).
76. Listgarten, J. & Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4**, 419–434 (2005).
77. Qian, W.-J., Jacobs, J.M., Liu, T., Camp, D.G. II & Smith, R.D. Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol. Cell. Proteomics* **5**, 1727–1744 (2006).
78. Anderson, L. & Hunter, C.L. Quantitative mass spectrometric MRM assays for major plasma proteins. *Mol. Cell. Proteomics* **5**, 573–588 (2006).
79. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol. [online]* **5**, R80 (2004).
80. Meng, F., Forbes, A.J., Miller, L.M. & Kelleher, N.L. Detection and localization of protein modifications by high resolution tandem mass spectrometry. *Mass Spectrom. Rev.* **24**, 126–134 (2005).
81. Han, X., Jin, M., Breuker, K. & McLafferty, F.W. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **314**, 109–112 (2006).
82. Chait, B.T. Chemistry: mass spectrometry: bottom-up or top-down? *Science* **314**, 65–66 (2006).
83. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
84. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
85. Perkins, D.N., Pappin, D.J.C., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
86. Clauser, K.R., Baker, P. & Burlingame, A.L. Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71**, 2871–2882 (1999).
87. Zhang, N., Aebersold, R. & Schwilkowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2**, 1406–1412 (2002).
88. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
89. Colinge, J., Masselot, A., Giron, M., Dessingy, T. & Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3**, 1454–1463 (2003).
90. Matthiesen, R., Trelle, M.B., Hojrup, P., Bunkenborg, J. & Jensen, O.N. VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.* **4**, 2338–2347 (2005).
91. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6**, 654–661 (2007).
92. Craig, R., Cortens, J.P. & Beavis, R.C. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **19**, 1844–1850 (2005).
93. Johnson, R.S. & Taylor, J.A. Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol.* **22**, 301–315 (2002).
94. Frank, A. & Pevzner, P. PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
95. Ma, B. *et al.* PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
96. Hernandez, P., Gras, R., Frey, J. & Appel, R.D. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* **3**, 870–878 (2003).
97. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol. [online]* **6**, R9 (2005).
98. Rauch, A. *et al.* Computational proteomics analysis system (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **5**, 112–121 (2006).
99. Martens, L. *et al.* PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537–3545 (2005).
100. Li, X.J., Zhang, H., Ranish, J.A. & Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **75**, 6648–6657 (2003).
101. MacCoss, M.J., Wu, C.C., Liu, H.B., Sadygov, R. & Yates, J.R. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.* **75**, 6912–6921 (2003).
102. Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica* **12**, 111–139 (2002).
103. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
104. Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
105. Fermin, D. *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol. [online]* **7**, R35 (2006).
106. Tanner, S. *et al.* Improving gene annotation using peptide mass spectrometry. *Genome Res.* **17**, 231–239 (2007).
107. Edwards, N.J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol. [online]* **3**, 102 (2007).
108. Pedrioli, P.G.A. *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459–1466 (2004).
109. Martens, L. *et al.* Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* **5**, 3501–3505 (2005).