Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS

Alexey I. Nesvizhskii and Ruedi Aebersold

Tandem mass spectrometry has been used increasingly for high-throughput analysis of complex protein samples. A major challenge lies in the consistent, objective and transparent analysis of the large amounts of data generated by such experiments and in their dissemination and publication. Here, we review currently available computational tools and discuss the need for statistical criteria in the analysis of large proteomics datasets.

Alexey I. Nesvizhskii*
Ruedi Aebersold
Institute for Systems Biology
1441 N 34th Street
Seattle
WA 98103, USA
*e-mail:
nesvi@systemsbiology.org

▼ An explicit goal of proteomics is the identification and (if applicable) quantification of proteins expressed in a cell or tissue [1]. Apart from emerging technologies such as protein chips [2] and the mass spectrometric (MS) identification of intact proteins [3] (top-down proteomics), proteomic studies frequently depend on MS analysis of peptides generated by proteolysis of single purified proteins or protein mixtures. Over the past few years, analysis of complex protein mixtures by tandem MS (MS–MS) has become widely used. In this method, complex protein mixtures are digested with proteases and the resulting peptide samples separated by one- or multi-dimensional liquid chromatography (LC) and analyzed by MS and MS–MS to sequence the peptides (Figure 1) [4]. If the peptides are also encoded with a stable isotope signature, relative protein abundance with respect to a control sample can be accurately determined using the same platform [5,6]. Each MS–MS spectrum is associated with the amino acid sequence it best represents and the data obtained from all the spectra in an experiment are then used to infer the identity and quantity of proteins in a sample mixture. In a typical experiment of this type, thousands of MS and MS–MS spectra are generated. Sequence database searching of MS–MS spectra to determine the sequence of the precursor peptide is typically the first and often the only analysis carried out with such data. It is being increasingly recognized that more extensive analysis of proteomic data generated by LC–MS–MS experiments is required if the results generated from different experiments, instruments and laboratories are to be published and related to each other [7–9]. Here, we discuss the need for statistical criteria for the consistent analysis of large proteomics datasets and summarize currently available bioinformatics tools that support data analysis and processing in high-throughput proteomics-based LC–MS–MS.

Peptide identification

Analysis of proteomics datasets generated using LC–MS–MS usually starts with identification of the peptides that produce the acquired MS–MS spectra. In high-throughput studies, peptides are usually identified by sequence database searching of uninterpreted MS–MS spectra, and several algorithms have been developed for this purpose [10–21] (Table 1). In this approach, each acquired MS–MS spectrum is compared with theoretical spectra obtained from a sequence database. The search is typically restricted to only those database peptides that have a calculated mass within a small range of the measured peptide mass. Expectation of certain peptide properties with regard to the proteolytic enzyme specificity can be used as an additional

1359-6446/04/$ – see front matter PII: S1359-6446(03)02978-7

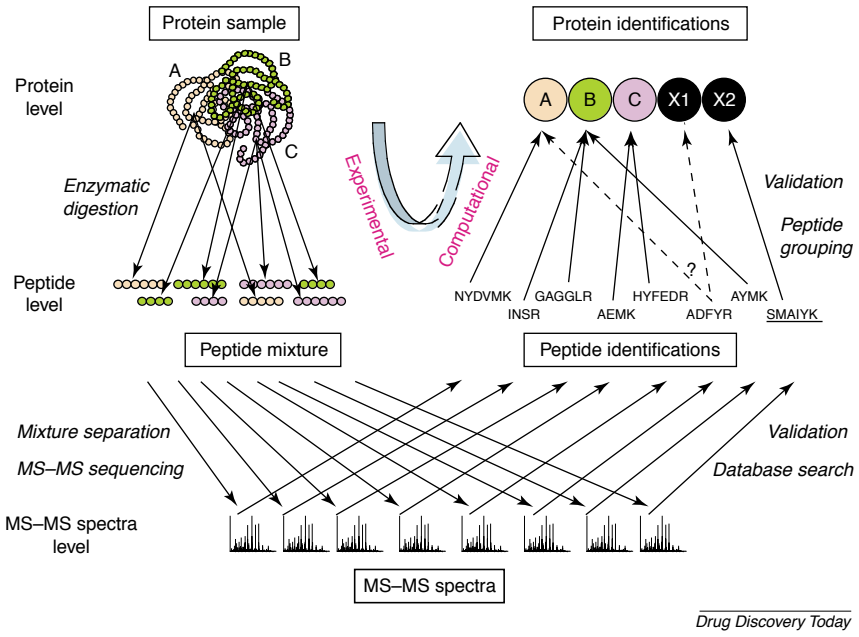

Figure 1. Experimental procedures and flow of data in a typical analysis of a complex protein mixture based on high-throughput liquid chromatography (LC) coupled with tandem mass spectrometry (MS–MS). First, sample proteins are proteolytically cleaved into small peptides. After separation using one- or multi-dimensional chromatography, peptides are ionized and selected ions fragmented to produce MS–MS spectra. Computational tools are used assign a peptide to each acquired MS–MS spectrum. The next step is to determine which proteins are present in the original sample (A, B and C) and which are false identifications (X1 and X2) corresponding to incorrect peptide assignments (underlined peptide sequence). The process of inferring protein identities is complicated by the presence of peptides corresponding to more than a single entry in the protein sequence database (indicated by a question mark), which can also lead to incorrect protein identifications.

constraint during the search. Theoretical spectra are calculated for each of the candidate peptides using common peptide fragmentation rules and then the theoretical and acquired MS–MS spectra are compared. Each acquired MS–MS spectrum is thereby assigned the best matching database peptide.

The main difference between database search programs is the scoring function used to quantify the degree of similarity between the compared spectra. Because the method involves the identification of peptides and not proteins, all types of sequence databases can be searched. These include protein sequence databases (most commonly searched) as well as genomic and expressed sequence tag (EST) databases [22,23]. However, it should be noted that the database search approach (in its straightforward use) only enables identification of those peptides that are present in the searched sequence database. It cannot therefore identify peptides derived from post-translationally modified proteins, sequence variants of known proteins, or proteins from partially sequenced genomes. Searching genomic or EST databases can potentially lead to the identification of novel proteins or novel splice variants of known proteins. However, owing to their size, searching such databases often takes a significant amount of time. In addition, other factors such as frame-shifts, incorrectly predicted open reading frames and the poor quality of many EST sequences further complicate the search. In the future, more-refined sequence databases from ongoing bioinformatics efforts, such as the Alternative Splicing Annotation Project [24] and the Alternative Splicing Database Project (http://www.ebi.ac.uk/asd/), might eliminate the need for searching MS–MS data against genomic databases.

An alternative approach to peptide identification is to determine peptide sequences from MS–MS spectra directly using *de novo* sequencing algorithms [25–33]. Derived peptide sequences can then be searched against a protein sequence database using BLAST- or FASTA-type sequence similarity search algorithms to infer the identities of their corresponding proteins [12,25,34]. However, currently available *de novo* sequencing programs are computationally intensive and require high quality MS–MS data. Such programs are therefore rarely used in high-throughput studies. They are typically used after database searching and applied only to a subset of the acquired data (high quality MS–MS spectra that did not get assigned a peptide with high confidence using the database search approach). In addition to being a method of primary peptide identification, *de novo* sequencing algorithms can also be used to simply filter out low-quality spectra or assist in validation of peptide assignments made by the database search tools [25]. Hybrid approaches have been described that combine the inference of short sequence tags (partial sequences) from MS–MS spectra using *de novo* sequencing-like algorithms with an error-tolerant database search (i.e. a search that allows for one-or-more mismatches between the peptide represented by the MS–MS spectra and the database sequence) [35–37]. A somewhat different strategy that also involves extraction of sequence tags has been recently proposed [18]. Another interesting approach is based on pattern recognition of peptide sequence motifs in MS–MS spectra [38]. It is hoped that these

Table 1. Publicly available tools for assigning peptides to tandem mass spectrometry spectra and for statistical validation of peptide and protein identifications

Program	Refs	Website
Database search tools		
SEQUEST	[10]	http://www.thermo.com
MASCOT	[11]	http://www.matrixscience.com[a]
MS-Tag	[12]	http://prospector.ucsf.edu[a]
Sonar	[13]	http://65.219.84.5/service/prowl/sonar.html[a]
ProbID	[21]	http://projects.systemsbiology.net/probid[b]
X! tandem	[26]	http://www.proteome.ca/opensource.html[b]
XProteo		http://xproteo.com:2698[a]
***De novo* sequencing tools**		
Lutefisk	[25]	http://www.hairyfatguy.com/Lutefisk[b]
De Novo	[34]	http://hto-c.usc.edu:8000/msms/menu/denovo.htm[a]
PEAKS	[5]	http://www.bioinformaticssolutions.com/Software/peaks/index.php[a]
Sequence tag approach		
GutenTag	[37]	http://fields.scripps.edu/GutenTag/
Integrated proteomics platform (multiple tools)		
SpectrumMill		http://www.chem.agilent.com/
Statistical validation of peptide and protein identifications		
PeptideProphet	[39]	http://www.proteomecenter.org/software.php[b]
ProteinProphet	[40]	http://www.proteomecenter.org/software.php[b]

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

methods and the error-tolerant database search approach [20] will eventually lead to the development of publicly available computational tools for the identification of post-translationally modified or mutated peptides that can be automated for use in a high-throughput environment. Nevertheless, direct database searching will probably continue to be used as the primary peptide identification method in most high-throughput LC–MS–MS-based studies.

Validation of peptides identified by database searching

As with many other database search applications, the main challenge is not finding the best match in the database but rather how to determine whether this best match assignment is correct [39–41]. If all spectra acquired in a typical LC–MS–MS experiment are searched against a sequence database, and the best match is assumed to be correct, then (without further filtering) a large fraction of the assigned peptides would be wrong [42]. This situation can arise because the scoring schemes used in current database search tools are based on a simplified representation of the peptide ion fragmentation process. In addition, the charge state of the peptide ions selected for fragmentation is not always known with high certainty, and many of the MS–MS spectra are of low quality. Furthermore, a significant number of high-quality spectra are assigned a wrong peptide because their true corresponding peptides are not present in the searched sequence database.

The sensitivity and specificity of the peptide identification process can be increased by several methods, including additional processing of MS–MS spectra before database searching [43], clustering of redundant spectra [44,45], removal of low-quality spectra [46,47], and application of automated charge-state determination algorithms [48–50]. In addition, the development of more-advanced scoring schemes that incorporate additional knowledge of peptide fragmentation chemistry [51,52] should result in further improvements. Nevertheless, the problem of incorrect peptide assignments can be only reduced and not completely eliminated. Therefore, to derive meaningful information from the data, significant effort has to be put into validation of peptide assignments produced by the database search tools [7,8,39]. This applies not only to tools that themselves provide no statistically computed confidence measures for evaluation of the validity of peptide identifications (i.e. SEQUEST [10]), but also to probability-based scoring tools such as MASCOT [11]. Manual validation of database search results is time-consuming and simply not

feasible for high-throughput analysis of large datasets containing hundreds of thousands of spectra. Furthermore, manual validation requires significant expertise in MS and peptide fragmentation chemistry, which is often not available, and consistent and objective evaluation of the data are difficult, even by experts.

As an alternative approach, or used in combination with manual validation, researchers can separate correct from incorrect peptide assignments by applying *ad hoc* filtering criteria based upon database search scores and some properties of the assigned peptides. This task can be facilitated by software tools such as INTERACT [53], DTASelect [54] or CHOMPER [55], all of which are compatible with both SEQUEST and MASCOT, the two most commonly used database search tools. However, with few exceptions [42,56], false identification error rates resulting from the application of filtering criteria are not estimated and not reported. Therefore, comparison of results from different experiments or groups is virtually impossible. Such comparisons are further complicated by the use of different database search tools for peptide assignment. Thus, consistent and reliable interpretation of data to enable the comparison of results from different experimental groups will require robust statistical methods to validate peptide assignments to MS–MS spectra. Similar advantages have been already realized in other high-throughput fields. For example, statistical models have been developed for the estimation of errors in raw DNA sequences obtained using large-scale DNA sequencing [57].

Several statistical methods for validating peptide assignments to MS–MS spectra made by database search tools have recently been described [39,58–61]. Fenyo and Beavis [60] converted the scores reported by database search tools into expectation values similar to those used in the sequence similarity search algorithms [62]. If such an approach were universally accepted, the problem of incompatibility between scoring schemes in different search tools would be eliminated. However, expectation values do not enable the estimation of false-positive error rates resulting from filtering of the data. Furthermore, this approach requires significant modification of already existing tools, which is unlikely to occur. Instead, peptide assignments can be validated using statistical programs developed on top of existing database search tools. This approach has the additional advantage that, in principle, it can use any additional information that discriminates between correct and incorrect peptide assignments but is not used as a part of the database search scoring scheme. Such additional information can be obtained directly from the sequences of assigned peptides, for example, the observed frequency of missed cleavage sites in the peptides sequences (internal residues at which the protein was expected to be cleaved by the proteolytic enzyme), or the separation coordinates of a peptide based on reverse-phase elution time (reverse-phase chromatography) [63] or *pI* value (isoelectric focusing gels) [64]. Therefore, the peptide identification process can itself be assisted by knowledge of the protein digestion and peptide separation processes.

Several supervised classification methods for post-database search validation of peptide assignments have recently been described, with underlying statistical methods based on linear discriminant analysis [39], support vector machines [59] and non-linear function optimization [61]. However, it should be noted that fully supervised classification algorithms might not produce accurate results when applied to datasets that are significantly different from those used for training. This limits the realistic applicability of such methods in a high-throughput environment owing to variations in the quality of acquired MS–MS spectra, complexity of the analyzed samples, and differences in the experimental protocols, among other factors. Thus, rather than being relied upon exclusively, training datasets should be used to determine the features that discriminate between correct and incorrect peptide assignments from the data itself [39].

One statistical model used in the software tool PeptideProphet is based upon use of the expectation maximization algorithm to derive a mixture model of correct and incorrect peptide assignments from the data [39]. It uses the observed information about each assigned peptide in the dataset, learns to distinguish correct from incorrect peptide assignments and, finally, computes a probability for each assignment being correct. Peptide assignment information used by the model typically includes database search scores, the difference between measured and theoretical peptide mass, the number of termini consistent with the type of enzymatic cleavage used, and the number of missed cleavage sites. If the database search tool outputs more than a single score useful for distinguishing correct from incorrect peptide assignments, all such scores are combined into a single score (discriminant score) in such a way that correct and incorrect peptide assignments are optimally discriminated for in every type of mass spectrometer. The model also uses additional information where available, such as the presence of a specific amino acid or sequence motif. For example, the presence of cysteine confers avidin-affinity-purification of peptides containing biotinylated cysteines [41], and the sequence motif N-X-S/T discriminates peptides containing *N*-linked glycosylation sites [65]. This can also be extended to include peptide separation coordinates such as elution time [63] or *pI* value [64]. Finally, because this method learns from the data,

it can handle variations in sample complexity, data quality and proteolytic digest efficiency, among other factors.

The advantage of using probabilities as confidence measures to accompany peptide identifications is that they can be used to estimate both the total number of correct identifications and the false-positive error rates resulting from data filtering using a minimum computed probability as the filtering criteria [39]. This facilitates the comparison of different types of mass spectrometers or the benchmarking of various mass spectrometer settings and experimental procedures to identify those that maximize the number of correct peptide identifications per sample or per unit time. More importantly, computed peptide probabilities enable statistical estimations of the presence of proteins that correspond to those peptides in the original sample.

Validation of protein identification

The goal of a high-throughput proteomics approach is to determine the identity of the proteins present in the original sample. However, because MS–MS spectra are produced from peptides and not proteins, all conclusions drawn about the protein content of the original sample are based upon the identification of peptides. The connectivity between peptides and proteins is usually quite straightforward when based on the digestion of purified proteins. This is the case in most studies in which proteins extracted from 2D gels are analyzed. However, this connectivity is lost when complex protein samples are digested [40,66]. As a result, inferring protein identities from the set of identified peptides becomes a major challenge (Figure 1). As a first step, peptide assignments must be grouped according to their corresponding entries in the protein sequence database. Next, for each

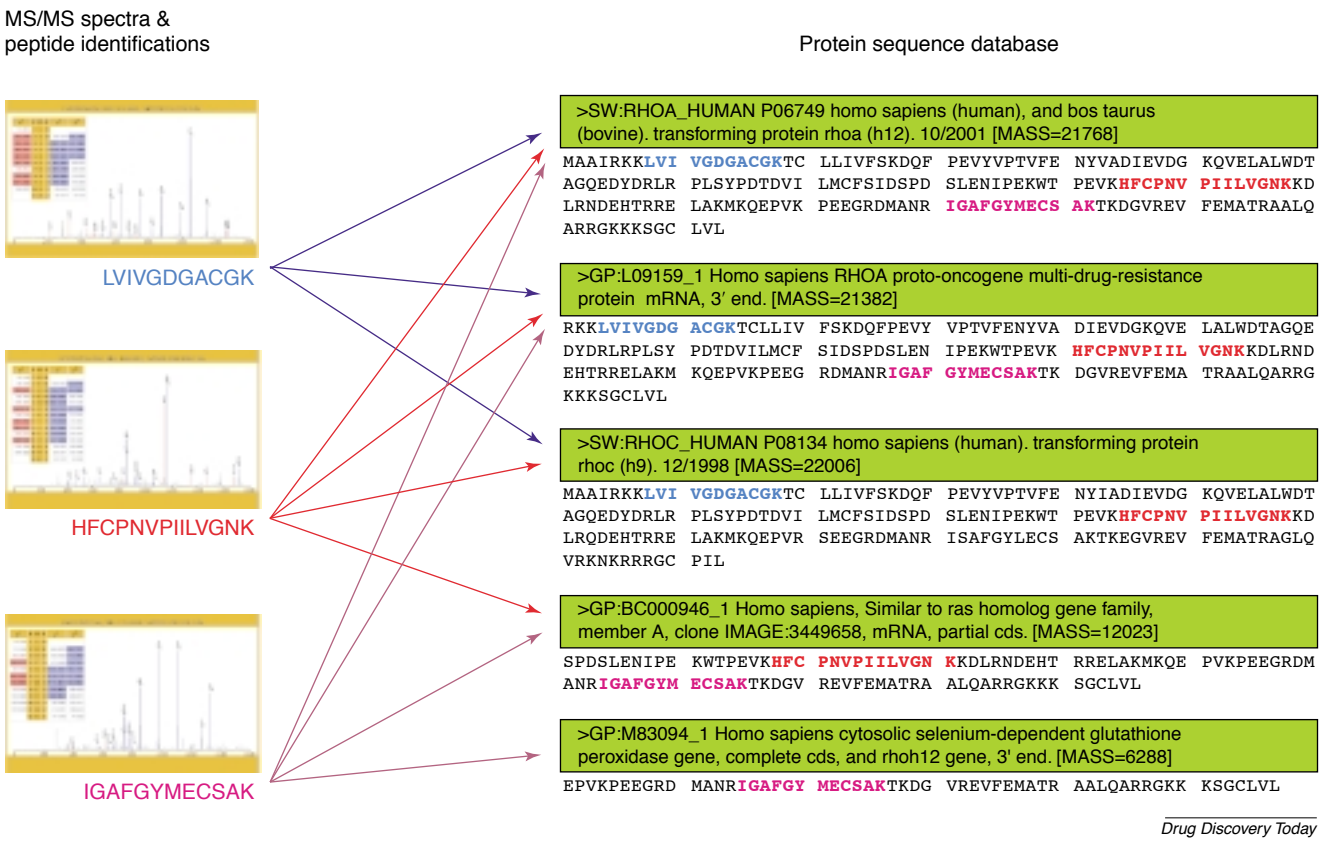

Figure 2. Identification of peptides whose sequence is present in more than one entry in the protein sequence database makes it difficult to infer the particular corresponding protein(s) in the sample. In this example (dataset from Ref. [41]), a total of three peptides were identified corresponding to five entries in the searched protein sequences database. There are no peptides corresponding to only one of the entries in the group and not another. Manual examination reveals that three of the five entries are redundant (partial sequences of the RHOA_HUMAN protein) and should be excluded from consideration. Between the remaining two entries, RHOA_HUMAN corresponds to all three peptides, whereas RHOC_HUMAN corresponds to only two peptides. RHOA_HUMAN is therefore the most probable candidate as its presence in the sample is necessary to explain all observed peptides. The data do not provide conclusive evidence for the presence of RHOC_HUMAN in the sample.

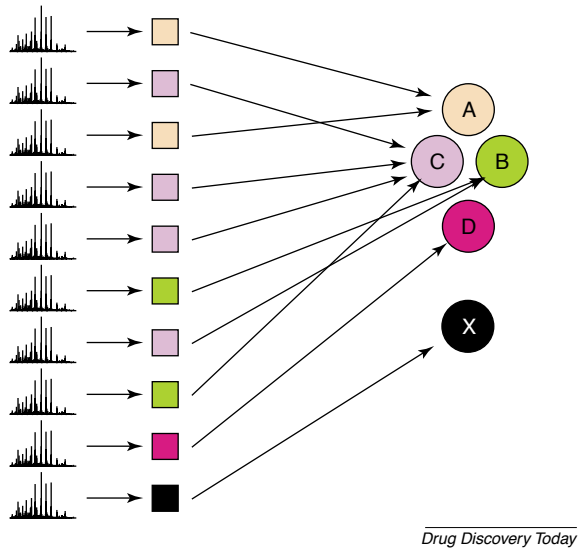

Figure 3. Non-random grouping of peptides according to their corresponding proteins. Ten tandem mass spectrometry (MS–MS) spectra were searched against a protein sequence database and each spectrum was assigned the best matching peptide with nine out of ten assignments being correct. The incorrect peptide assignment (black square) results in one incorrect protein identification (X). Nine correct peptide assignments (colored squares) correspond to only four correct proteins (A, B, C and D). As a result, in this example, a 10% false-positive identification error rate at the peptide level (nine out of ten peptide assignments are correct) translates into a 20% error rate at the protein level (four out of five protein identifications are correct).

protein, the combined peptide evidence is used to estimate the likelihood of its presence in the sample.

Assembling peptides into proteins is not straightforward (Figure 2). This challenge is analogous to that of shotgun fragment assembly where overlapping short DNA segments must be ordered to recreate the original sequence [67]. Determining the correct sequence assembly is difficult owing to the presence of repeats (identical stretches of the DNA sequence present at different locations throughout the genome). Similarly, the presence of degenerate peptides, that is, peptides whose sequence is present in more than one entry in the protein sequence database, makes it difficult to determine the corresponding protein(s) present in the sample [40]. Such cases often result from the presence of homologous proteins, splicing variants, or redundant entries in the protein sequence database, and are particularly abundant in large higher eukaryotic databases [66]. Unfortunately, this problem is often overlooked, and the words 'protein identification' and 'peptide identification' are used almost interchangeably.

Another challenge arises from the non-random grouping of peptides according to their corresponding proteins (Figure 3) [40]. Correct peptide identifications tend to group into a relatively small number of proteins. By contrast, incorrect peptide assignments can be described as random matches to entries in a very large protein sequence database. Thus, almost every (high scoring) incorrect peptide assignment results in one additional incorrect protein identification. As a result, even a relatively small false-positive identification error rate at the peptide level can translate into a significant error rate at the protein level. It also makes detection of correct protein identifications based on a single peptide (often the case with low abundance proteins) difficult because most of the incorrect protein identifications only have one corresponding peptide in the dataset.

Some database search tools (e.g. MASCOT) enable the user to view the results in a format that groups peptides according to their corresponding proteins. However, most large-scale studies generate multiple datasets of MS–MS spectra that are acquired and processed at different times. Thus, to derive a composite list of protein identifications, peptide assignments from multiple experiments must be combined using other means. The software tools INTERACT, DTASelect and CHOMPER can be used to automate this process. However, these tools do not compute any statistical confidence measures for protein identifications. Other recently described programs compute some kind of probability-based scores [58,61,68] but do not use any statistical models for resolving degenerate peptides.

The statistical model of Nesvizhskii *et al.* [40] used in the software tool ProteinProphet addresses all of the difficulties discussed previously. It computes a probability that a protein is present in the sample by combining the probabilities that corresponding peptides are correct. Individual peptide probabilities are adjusted for observed protein grouping information. Peptides corresponding to single-hit proteins are penalized (but not excluded), whereas those corresponding to multi-hit proteins are rewarded. The amount of adjustment depends on the sample complexity and the number of acquired MS–MS spectra, among other factors, and is learned from the data using the expectation maximization algorithm. The model handles degenerate peptides by sharing each such peptide among all its corresponding proteins to derive a minimal protein list sufficient to account for the identified peptides. The model reduces redundant database entries into a single identification and groups together those proteins that are impossible to differentiate on the basis of identified peptides. The model produces accurate probabilities of the presence of a protein, with high power to

discriminate between correct and incorrect protein identifications including identifications based on a single peptide. Furthermore, computed probabilities can be used to estimate false-positive error rates resulting from data filtering. This model therefore provides a consistent means of publishing large-scale datasets of protein identifications [40,41].

Large-scale datasets: filtering and their publication in the literature

Computational tools for the statistical validation of peptide and protein identifications are of significant value to high-throughput proteomics. They avoid laborious manual data validation and provide a fast, consistent and transparent means to analyze data. As these tools become widely available, and independently tested and understood by researchers collecting and analyzing the data, they could provide a standard for the publication and dissemination of large-scale protein identification datasets [8,40,41]. When only the most confident identifications are desired, such as in the submission of protein identification data to (not yet existing) public databases, datasets of protein identifications can be filtered using a high minimum-probability threshold (e.g. 0.99). Researchers publishing large-scale datasets in scientific journals should be encouraged to include extended lists of peptide and protein identifications (e.g. all identifications with a >0.5 probability of being correct) along with their corresponding probabilities [41]. Ideally, publications should also include all supporting data, including MS–MS spectra, although the practical aspects of storing and managing large datasets by scientific journals have yet to be worked out. If this protocol is followed, other researches will have access to the most complete dataset possible to interpret or use further at their discretion. For example, researchers accessing the published data might be interested in a particular set of proteins, regardless of the statistically estimated level of confidence for their presence in the sample. In such cases, the raw MS–MS data can be further interrogated using additional computational approaches such as those based on *de novo* sequencing. Often, additional experiments will be necessary to confirm the validity of some protein identifications. However, time and cost considerations mean that such experiments might only be possible for a small number of proteins. Protein probabilities can therefore serve as a guide for selection of the most interesting candidates.

Computed protein probabilities should also enable the user to compare different protein identification datasets (e.g. those generated by different research groups studying the same biological system) using the total number of correct protein identifications estimated by the model [39–41]. Datasets can also be compared objectively by specifying a uniform error rate and applying to each dataset the corresponding minimum probability threshold as the data filter. Finally, published protein identifications accompanied by accurate probabilities will provide maximal information to higher level computational analyses based on proteomic data, such as those concerned with the identification of protein–protein interactions or metabolic pathway reconstruction, provided that protein probabilities are taken into account.

Concluding remarks

Significant progress in protein chemistry, separation methods and mass spectrometry in the past decade has enabled datasets containing information about thousands of proteins to be collected in a matter of weeks or even days. However, the development of computational tools for validation, interpretation and extraction of biological knowledge from such datasets has lagged behind. The process of validating datasets of protein identifications obtained using high-throughput LC–MS–MS traditionally relied upon time-consuming and often subjective manual verification. In the absence of data analysis standards, published large-scale datasets are of little value to other researchers; even the value of the interpreted data is unclear. Researchers interrogating such datasets cannot easily compare or correlate the findings with those of their own. Fortunately, the importance of data analysis using robust statistical criteria is now being realized. Several computational tools have been developed that enable the fast, consistent and transparent analysis of large-scale proteomics datasets.

Clearly, it is unrealistic in the short term to expect different research groups to agree to use the same statistical tools or approaches to analyze their data. However, it is reasonable for the general research community, particularly members of the editorial boards of scientific journals, to request that researchers statistically validate their data using only those tools that satisfy the following criteria. First, important details of the underlying statistical models should be made available. Second, accuracy of the models should be tested extensively using reference datasets (which can be created specifically for that purpose). Finally, the software tools used should be made available to the public. This could be achieved on a commercial basis for a reasonable fee or, where possible, free-of-charge and as open source programs. If widely accepted, together with the development of MS data representation standards [69], such an approach should facilitate the creation of centralized databases of peptide and protein identifications and public repositories for storing acquired MS data. The idea

of merging data from the same organisms – particularly humans – generated in different experiments is particularly attractive. The combined results could be then applied to the whole genome – eventually validating all genes that are expressed on the protein level – or used to elucidate global patterns (e.g. tissue specificity) of protein expression that would otherwise be missed in analysis of a single experiment. Finally, access to data stored in public repositories, especially raw data, will enable those with access to no MS data to become involved in the development of more advanced computational methods and software tools.

Acknowledgements

This work was funded in part by Federal funds from the National Heart, Lung and Blood Institute at the National Institutes of Health (contract number N01-HV-28179).

References

1 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
2 Zhu, H. and Snyder, M. (2003) Protein chip technology. *Curr. Opin. Chem. Biol.* 7, 55–63
3 Reid, G.E. and McLuckey, S.A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* 37, 663–675
4 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682
5 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
6 Goshe, M.B. and Smith, R.D. (2003) Stable isotope-coded proteomic mass spectrometry. *Curr. Opin. Biotechnol.* 14, 101–109
7 Burlingame, A.L. (2003) Toward deciphering the knowledge encrypted in large datasets. *Mol. Cell. Proteomics* 2, 425
8 Patterson, S.D. (2003) Data analysis – the Achilles heel of proteomics. *Nat. Biotechnol.* 21, 221–222
9 Boguski, M.S. and McIntosh, M.W. (2003) Biomedical informatics for proteomics. *Nature* 422, 233–237
10 Eng, J.K. *et al.* (1994) 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
11 Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567
12 Clauser, K.R. *et al.* (1999) Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71, 2871–2882
13 Field, H.I. *et al.* (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. *Proteomics* 2, 36–47
14 Bafna, V. and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 17 (Suppl.), S13–S21
15 Pevzner, P.A. *et al.* (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* 11, 290–299
16 Zhang, N. *et al.* (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2, 1406–1412
17 Havilio, M. *et al.* (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* 75, 435–444
18 Hernandez, P. *et al.* (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* 3, 870–878
19 Colinge, J. *et al.* (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* 3, 1454–1463
20 Sadygov, R.G. and Yates, J.R., III (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* 75, 3792–3798
21 Craig, R. and Beavis, R.C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 17, 2310–2316
22 Kuster, B. *et al.* (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 1, 641–650
23 Choudhary, J.S. *et al.* (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* 1, 651–667
24 Lee, C. *et al.* (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res.* 31, 101–105
25 Taylor, J.A. and Johnson, R.C. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 73, 2594–2604
26 Dancik, V. *et al.* (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 6, 327–342
27 Fernandez-de-Cossio, J. *et al.* (2000) Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for *de novo* sequencing by tandem mass spectrometry. *Electrophoresis* 21, 1694–1699
28 Chen, T. *et al.* (2001) A dynamic programming approach to *de novo* sequencing via tandem mass spectrometry. *J. Comput. Biol.* 8, 325–337
29 Lubeck, O. *et al.* (2002) New computational approaches for *de novo* peptide sequencing from MS/MS experiments. *Proc. IEEE* 90, 1868–1874
30 Lu, B.W. and Chen, T. (2003) A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 10, 1–12
31 Ma, B. *et al.* (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17, 2337–2342
32 Bafna, V. and Edwards, N. (2003) On *de novo* interpretation of peptide sequencing via tandem mass spectrometry. In *RECOMB 2003*, pp. 9–18
33 Cannon, W.R. and Jarmal, K.D. (2003) Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Commun. Mass Spectrom.* 17, 1793–1801
34 Shevchenko, A. *et al.* (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 73, 1917–1926
35 Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399
36 Sunyaev, S. *et al.* (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* 75, 1307–1315
37 Tabb, D.L. *et al.* (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* 75, 6415–6421
38 Liebler, D.C. *et al.* (2002) Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal. Chem.* 74, 203–210
39 Keller, A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74, 5383–5392
40 Nesvizhskii, A.I. *et al.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75, 4646–4658
41 Von Haller, P.D. *et al.* (2003) The application of new software tools to quantitative protein profiling via 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* 2, 428–442

42 Keller, A. *et al.* (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* 6, 207–212
43 Gentzel, M. *et al.* (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* 3, 1597–1610
44 Beer, I. *et al.* (2003) Pep-Miner: high throughput proteomics made easy. In *Proc. 50th ASMS Conf. Mass Spectrom. Allied Top.*, Montreal, Canada
45 Tabb, D.L. *et al.* (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.* 75, 2470–2477
46 Moore, R.E. *et al.* (2000) Method for screening peptide fragment ion mass spectra prior to database searching. *J. Am. Soc. Mass Spectrom.* 11, 422–426
47 Kolker, E. *et al.* (2003) Initial proteome analysis of model microorganism *Haemophilus influenzae* Rd strain KW20. *J. Bacteriol.* 185, 4593–4602
48 Sadygov, R.G. *et al.* (2002) Code developments to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome Res.* 1, 211–215
49 Perez, R.E. *et al.* (2002) Peptide precursor charge state determination directly from ion trap MS/MS spectra. In *Proc. 50th ASMS Conf. Mass Spectrom. Allied Top.*, Orlando, FL, USA
50 Colinge, J. *et al.* (2003) Improved peptide charge state assignment. *Proteomics* 3, 1434–1440
51 Tabb, D.L. *et al.* (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* 75, 1155–1163
52 Kapp, E.A. *et al.* (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* 75, 6251–6254
53 Han, D.K. *et al.* (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* 19, 946–951
54 Tabb, D.L. *et al.* (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 1, 21–26
55 Eddes, J.S. *et al.* (2002) CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. *Proteomics* 2, 1097–1103
56 Peng, J. *et al.* (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: the yeast proteome. *J. Proteome Res.* 2, 43–50
57 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194
58 MacCoss, M.J. *et al.* (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* 74, 5593–5599
59 Anderson, D.C. *et al.* (2003) 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
60 Fenyo, D. and Beavis, R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 75, 768–774
61 Kislinger, T. *et al.* (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* 2, 96–106
62 Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
63 Petritis, K. *et al.* (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* 75, 1039–1048
64 Cargile, B.J. *et al.* Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* (in press)
65 Zhang, H. *et al.* (2003) Identification and quantification of *N*-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* 21, 660–666
66 Rappsilber, J. and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* 27, 74–78
67 Myers, E.W. *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* 287, 2196–2204
68 Moore, R.E. *et al.* (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* 13, 378–386
69 Orchard, S. *et al.* (2003) Further advances in the development of a data interchange standard for proteomics data. *Proteomics* 3, 2065–2066