Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book

Rovshan G Sadygov, Daniel Cociorva & John R Yates III

Database searching is an essential element of large-scale proteomics. Because these methods are widely used, it is important to understand the rationale of the algorithms. Most algorithms are based on concepts first developed in SEQUEST and PeptideSearch. Four basic approaches are used to determine a match between a spectrum and sequence: descriptive, interpretative, stochastic and probability–based matching. We review the basic concepts used by most search algorithms, the computational modeling of peptide identification and current challenges and limitations of this approach for protein identification.

An unintended consequence of whole-genome sequencing has been the birth of large-scale proteomics. What drives proteomics is the ability to use mass spectrometry data of peptides as an 'address' or 'zip code' to locate proteins in sequence databases. Two mass spectrometry methods are used to identify proteins by database search methods. The first method uses a molecular weight fingerprint measured from a protein digested with a site-specific protease[1–5]. A second method uses tandem mass spectra derived from individual peptides of a digested protein[6,7] (**Fig. 1**). Because each tandem mass spectrum represents an independent and verifiable piece of data, this approach to database searching has the ability to identify proteins in mixtures, enabling a rapid and comprehensive approach for the analysis of protein complexes and other complicated mixtures of proteins[6,8–12]. New biology has been discovered based on fast and accurate protein identification[13–18]. As tandem mass spectral protein identification has proliferated, it has become increasingly important to understand the rationale of individual database search algorithms, their relative strengths and weaknesses, and the mathematics used to match sequence to spectrum.

In this review we discuss the prevailing fragmentation models, spectral preprocessing, methods to match tandem mass spectra to sequences and several approaches to matching tandem mass spectra of peptides whose exact sequences may not be present in the database. Space limitations restrict a detailed description of all algorithms in this rapidly expanding field. Also, some algorithms are proprietary, and thus, details on how they work are unknown. This review should supplement and update earlier reviews on database search algorithms[19–24].

Peptide fragmentation and data preprocessing

In tandem mass spectrometry (MS/MS), gas phase peptide ions undergo collision-induced dissociation (CID) with molecules of an inert gas such as helium or argon[25]. Other methods of dissociation have been developed, such as electron capture dissociation (ECD), surface induced dissociation (SID) and electron transfer dissociation (ETD), but gas-phase CID is the most widely used in commercial tandem mass spectrometers. The dissociation pathways are strongly dependent on the collision energy, but the vast majority of instruments use low-energy CID (<100 eV)[26]. At low collision energies fragmentation mainly occurs along the peptide backbone bonds, whereas at higher energies fragments generated by amino acid side-chains are observed[25,27]. At low-energy CID, conditions normally used in triple quadrupole, quadrupole

Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, USA. Correspondence should be addressed to J.R.Y. (jyates@scripps.edu).

Figure 1 | Overview of the protein identification process. A protein mixture is digested, and the resulting peptides are analyzed by MS/MS to obtain experimental spectra. Search programs find database candidate sequences whose theoretical spectra are compared to the experimental spectrum. The best match (highest-scoring candidate sequence) defines the identified database peptide and the corresponding database protein. Validation software then determines whether the peptide and protein identifications are true or false.
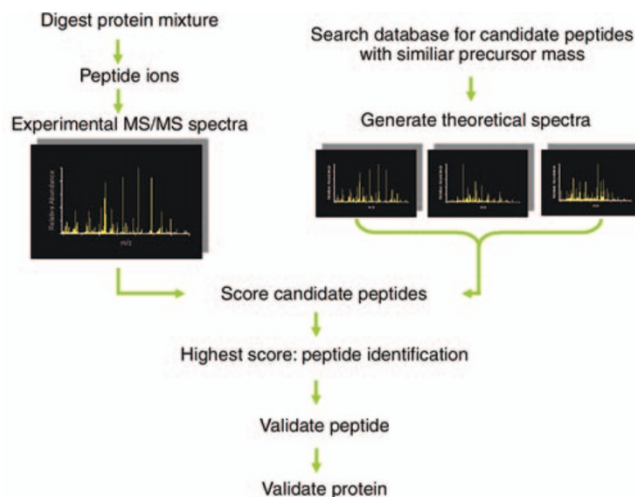
time of flight and ion trap (both linear and Paul) mass spectrometers, b-ions, y-ions and neutral losses of water and ammonia dominate the mass spectrum (**Box 1**). Fragmentation patterns are also strongly dependent on the chemical and physical properties of the amino acids and sequences of the peptide[28,29]. Most algorithms assume that peptides preferentially fragment into b- and y-ions. The distribution of intensities between b- and y-ions is the subject of intensive studies, and this distribution can vary by type of instrument (for example, ion trap as compared to Q-TOF), but this information is not yet fully exploited in most matching models. A mobile proton model[30] has been proposed to explain intensity patterns observed in MS/MS, and Zhang has developed a theoretical model that predicts fragment ion intensities well[31]. As with any measurement process, tandem mass spectra may have some level of uncertainty. The accuracy of the mass-to-charge ratio (m/z) and the mass resolving power are limited, electronic and chemical noise may be present and ion signals may fluctuate

BOX 1   PEPTIDE FRAGMENTATION

Low-energy CID dissociates the amide bond along the peptide backbone. As a result, two fragments are produced, one containing the N terminus and the other containing the C terminus. Nomenclature denotes the N-terminal fragments with letters a, b and c and the C-terminal fragments with letters x, y and z[59]. Internal ion fragments are formed by simultaneous cleavage of N and C termini. Immonium ions are of the structure HN = CH–R[25]. A numerical subscript for each fragment ion indicates the position of the amino acid at which the bond cleavage occurs. For N-terminal fragments, the numbering starts from the N terminus, and for C-terminal fragments, the numbering starts from the C terminus.

Fragment ion nomenclature. Schematic diagram of N-terminal a_2, b_2 and c_2 ions and C-terminal x_3, y_3 and z_3 ions for a five-amino-acid peptide.
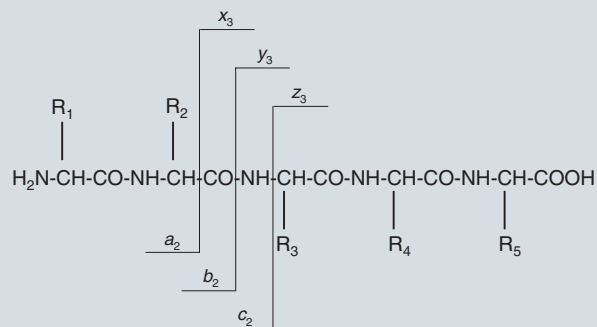
as a result of changes in the concentrations of peptides entering the ion source. Given a precursor m/z value and a list of fragment ions, the goal is to match these to an amino acid sequence within the measurement and fragmentation uncertainty of the mass spectrometer (**Fig. 1**).

Some of the methods described below are used in our laboratory and illustrate a general approach for the analysis of large data sets. One source of uncertainty in analyzing tandem mass spectra of multiply charged ions is determination of the charge state of the precursor peptide, which is critical to accurately calculating a peptide's molecular weight. A highly accurate molecular weight measurement or calculation can be very effective in restricting search results. Methods to determine the charge state of ions involve deconvolution or determination of the m/z spread of isotope peaks[32]. On low-resolution instruments, where these methods are ineffective, MS/MS of peptides with charge states higher than 1 are typically searched twice, once calculating a molecular weight assuming that the charge state is +2 and the second time assuming the charge state is +3. Based on observations by Dancik and colleagues[33] that complementary fragment ions (N-terminal and C-terminal fragment ions) can be used to improve molecular weight calculations, our group and others used a variation of this approach to determine peptide ion charge state[34,35]. In good-quality tandem mass spectra there are numerous complimentary ions; thus, if the precursor ion is assumed to be doubly charged, the complementary ions present in the spectrum should sum to this molecular weight. If the ion is triply charged, then the complementary fragment ions will sum to this molecular weight. Both situations are tested, and the molecular weight calculation that accounts for the most complementary ions is assumed to be the correct charge state. In addition to the above method, ion traps can also be used to perform a narrow mass range scan at resolutions sufficient to determine the charge state, but this necessitates an additional scan and reasonably abundant signal[36]. The newer linear ion traps with higher scan speeds can accommodate the high-resolution scan without decreasing the efficiency of data acquisition.

Most large-scale tandem mass spectrometry data is acquired using automated methods such as data-dependent data acquisition, which triggers MS/MS based on ion abundance. When the ion abundance level to trigger MS/MS is set just above the background noise level, MS/MS data is almost continuously acquired.

BOX 2 DATABASE SEARCHING ALGORITHMS

Descriptive models
Descriptive algorithms are based on a mechanistic prediction of how peptides fragment in a tandem mass spectrometer, which is then quantified to determine the quality of the match between the prediction and the experimental spectrum. Mathematical methods such as correlation analysis have been used to assess match quality.

Interpretative models
Interpretative approaches are based on manual or automated interpretation of a partial sequence from a tandem mass spectrum and incorporation of that sequence into a database search. Matches between the sequence and the spectrum have been scored using probabilities or correlation methods.

Stochastic models
Stochastic models are based on probability models for the generation of tandem mass spectra and the fragmentation of peptides. Basic probabilities of fragment ion matches are obtained from training sets of spectra of known sequence identity. Stochastic models use statistical limits on the measurement and fragmentation process to create a likelihood that the match is correct.

Statistical and probability models
Statistical and probability models determine the relationship between the tandem mass spectrum and sequences. The probability of peptide identification and its significance are then derived from the model.

Distinguishing data acquired from nonpeptide ions or identifying poor-quality MS/MS of peptides can potentially lower false-positive rates, as these spectra should not correctly match to peptide sequences, but most algorithms will still attempt a match to a sequence. Peptide ions selected for MS/MS at very low signal levels will produce spectra with poor signal-to-noise ratios and often incomplete sequence ions. Methods have been devised to sort the good from the bad spectra[34,37,38]. Recently, Bern *et al*. presented two different algorithms for assessing spectral quality prior to a database search: a binary classifier, which predicts whether or not the search engine will be able to make an identification, and a statistical regression, which predicts a more universal quality metric, independent of the database search program[39]. A quadratic discriminant analysis, a classical machine learning algorithm, was trained on a data set of manually validated good and bad spectra. Each spectrum is assigned a 'feature vector,' including number of peaks, total intensity and relative normalized intensity of the peaks that (i) differ by the mass of an amino acid, (ii) differ by a mass of 18 Da (mass of a water molecule) (iii) add up to the mass of the precursor ion and (iv) have associated isotope peaks. We report that the parameters with the best discriminant power are the relative normalized intensity of the peaks that differ by the mass of an amino acid and the relative normalized intensity of complimentary ions.

Review of database search algorithms
The goal of a tandem mass spectral database search is to identify the best sequence match to the spectrum. For tandem mass spectra with good signal-to-noise ratio and uniform fragmentation, it is reasonably straightforward to identify the correct sequence match. In situations where a tandem mass spectrum is of poorer quality or the peptide ion undergoes unusual fragmentation, an analysis may benefit from the use of multiple search algorithms. A number of algorithms and scoring models have been developed to assess the likelihood of a match. They can show different selectivity and sensitivity at the edge of good spectral quality, and some programs have enough flexibility to permit the use of different types of MS/MS data or modification patterns[6,7,40–53]. Four basic approaches have been developed to model matches to sequences: descriptive, interpretative, stochastic and probability-based modeling (**Box 2**). Some of these programs can be accessed through websites (**Table 1**), but most are run on local computers to allow large-scale analyses (**Box 3**).

Descriptive models for database searching
SEQUEST is an example of a program that uses a descriptive model for peptide fragmentation and correlative matching to a tandem mass spectrum[6]. It uses a two-tiered scoring scheme to assess the quality of the match between the spectrum and amino acid sequence from a database. The first score calculated, the preliminary score (S_p), is an empirically derived score that restricts the number of sequences analyzed in the correlation analysis. S_p sums the peak intensity of fragment ions matching the predicted sequence ions and accounts for the continuity of an ion series and the length of a peptide. The original S_p score is:

$$S_p = \left(\sum_k I_k\right) m(1+\beta)(1+\rho)/L$$

Table 1 Database search programs and websites where information about these programs can be obtained

Program	Web site
Mascot	http://www.matrixscience.com/
Masslynx	http://www.waters.com
MS-Tag/MS-Seq	http://prospector.ucsf.edu/
PeptideSearch	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html
PepFrag	http://prowl.rockefeller.edu/PROWL
ProbID	http://projects.systemsbiology.net/probid
SEQUEST	http://fields.scripps.edu or http://www.thermo.com
SpectrumMill	http://www.chem.agilent.com/
X!Tandem	http://www.thegpm.org/TANDEM/index.html

where the first term in the product is the sum of ion abundances of all matched peaks, m is the number of matches, β is a 'reward' for each consecutive match of an ion series (for example, 0.075), ρ is a 'reward' for the presence of an immonium ion (for example, 0.15) and L is the number of all theoretical ions of an amino acid sequence.

The second score is a cross-correlation of the experimental and theoretical spectra. This score is referred to as XCorr. The theoretical spectrum is generated from the predicted fragment ions, the b- and y-ions for each of the sequences. In the theoretical spectrum the main ion series products are assigned an abundance of 50, a window of 1 atomic mass unit around the main fragment ions is assigned intensity 25, and water and ammonia losses are assigned intensity of 10. The theoretical and normalized experimental spectra are cross-correlated to obtain similarities between the spectra. First, a cross-correlation of the two discrete data sets, experimental data (E) and theoretical spectrum (T), is taken:

$$Corr(E,T) = \sum_{I=0}^{N-1} x_i y_{i+\tau}$$

The correlation is processed and averaged to remove the periodic noise in the interval of (−75 to 75). In addition to the preliminary and cross-correlations scores, SEQUEST produces another important quantity, normalized difference of Xcorr values between the best sequence and each of the other sequences. This value, ΔC_n, is important in distinguishing the best match from other lower-scoring matches. That is, ΔC_n is useful to determine the uniqueness of the match. If a match is reasonably unique to a sequence, the ΔC_n value will be large (>0.1). XCorr is independent of database size and reflects the quality of the match between spectrum and sequence, whereas ΔC_n is database dependent and reflects the quality of the match relative to near misses.

The cross-correlation score is a sensitive measure. However, like other measures based on additive features, it is dependent on peptide mass, charge state and spectral quality. Thus it has been observed that larger peptides score higher than similar-quality smaller peptides. Very dense (potentially noisy) spectra can have high cross-correlation scores. To address these issues, a few modifications have been made to the cross-correlation score. To normalize XCorr for spectral noise and peptide size, the XCorr value is divided by auto-correlation of the experimental spectrum or by the square root of the products of auto-correlations of experimental and theoretical spectra[54,55]. A statistical confidence can then be readily derived from the normalized cross-correlation scores. SEQUEST has been shown to have good sensitivity and flexibility and is applicable to data generated by different types of mass spectrometers[56,57].

Other programs in this group include SONAR[52] and SALSA[49]. To determine the quality of the spectral match, SONAR uses a dot product of experimental and theoretical spectra. The dot product is the zero shift cross-correlation. SALSA scores the correspondence between the experimental fragment ion series and theoretical ion series regardless of their absolute position on the m/z axis. A virtual ruler is used with the relative separations of ions fixed and then superimposed on the experimental mass spectrum by aligning the first ion in the ion series to the fragment ion with the highest experimentally determined m/z. SALSA is adept at identifying peptides with unanticipated modifications or amino acids.
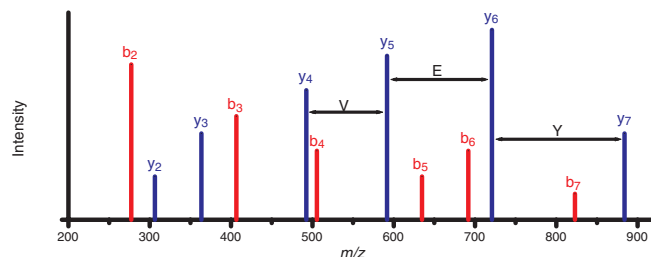
Figure 2 | Simplified representation of an MS/MS spectrum for the peptide IYEVEGMR. The b-ion ladder is shown in red and the y-ion ladder in blue. Distances between peaks on the horizontal mass-to-charge (m/z) axis can be used to infer partial sequences of the peptide. This example shows how the partial sequence YEV can be inferred from the y-ion ladder.

Interpretative models for database searching

PeptideSearch[7] is a program based on the assumption that in tandem mass spectra there is a continuous series of fragment ions that are clearly identifiable as a short amino acid sequence (**Fig. 2**) . A search engine has been fashioned using the partial sequence by dividing every candidate sequence into three parts: region 1 of unknown mass, region 2, containing the sequence tag and another region 3. The sequence ions associated with the sequence tag can be from the b- or y-ion series (defined in **Box 1**). Both possibilities are equally likely and must be considered by the algorithm. The assumption about the directionality of the partial sequence leads to the determination of masses of the region 1 (m_1) and 3 (m_3). For example, if the sequence ions are assumed to be b-ions, then m_1 is simply the mass-to-charge ratio of the smallest ion in the series, whereas m_3 is the difference between peptide mass and the mass-to-charge ratio of the largest ion in the series. The algorithm searches the database for sequences using the information from regions 1 (m_1), 2 (partial sequence tag) and 3 (m_3), as well as information from the peptide molecular weight, the protease specificity and mass accuracy. A sequence match is scored by computing the random probability match of each region and the N- and C-terminal amino acids expected from protease cleavage specificity. For the sequence tag it is assumed that the probabilities of all amino acids are equally likely. In this case, unique mass amino acids have a probability of 1/20, whereas amino acids with the same mass (within a specified accuracy) will have higher probability of $N \times 1/20$, where N is the number of amino acids with the same mass. Thus, for amino acids leucine and isoleucine, or glutamic acid and lysine, this probability will be 1/10, whereas for glycine it is 1/20. The probability of randomly matching regions 1 and 3 are equal and set to the ratio of mass accuracy to the average molecular weight of amino acid residues, 1/110 for a unit mass accuracy. Also, a probability is assigned to the amino acids at the cleavage sites. Complete tryptic cleavage of a protein results in peptides terminating in one of two amino acids—lysine or arginine. Therefore, if a sequence is tryptic, the random match probability is multiplied by 1/100, and if it is half tryptic by 1/10. Combining the probabilities of all regions and cleavage probabilities, a probability that a sequence match is random is set:

$$p_{random} = p_{NtermCleavage} \times p_{m_1} \times p_{1sttagposition} \times ... p_{lasttagposition} \times p_{m_3} \times p_{CtermCleavage}$$

The probability of a nonrandom match in a database with *N* amino acids would then be

$$p_{\text{nonrandom}} = (1 - 2 \times p_{\text{random}})^N$$

In the above formula the random match probability is multiplied by 2 to account for the fact that the direction of the partial sequence is not known. As it is seen from the formula for a nonrandom match, the identification is dependent on the size of the database. In general, the larger the database, the longer the sequence tag should be for higher confidence matches.

One of the advantages of the approach taken by PeptideSearch is that it is an error-tolerant algorithm. There could be a difference between the mass of the peptide as predicted in the database and the actual peptide. For example, the database sequence could have been predicted incorrectly, or there could have been a mutation or a post-translational modification in the peptide represented by the mass spectrum. The molecular weights of these altered peptides would not agree with molecular weights predicted from the database and they will not be identified by direct searches. PeptideSearch suggests using partial information from a combination of any two out of three regions, to identify peptides and map the region of alteration. If the origin of the sample is known, then the database search can be restricted to a smaller number of known proteins (for example, a particular species), in which case searches with short partial sequences (one or two amino acids) can yield reliable results.

The scoring model of PeptideSearch assumes the probability of each amino acid occurring is the same, 1/20. It is known, however, that amino acids occur in a database at different frequencies; for example, leucine occurs eight times more frequently than tryptophan). Furthermore, the probability scores are calculated using an incomplete model. The sum of probabilities of all database sequences is not 1: the probabilities for mass accuracy and enzyme specificity are arbitrary, as they represent a different probability space, and longer sequences naturally produce smaller probabilities. Despite these problems with the scoring model, sequence tagging is a useful approach for database searching.

Other implementations of the PeptideSearch algorithm are MS-Seq, by Clauser *et al.*, and the recently developed program GutenTag[42,47]. In GutenTag, the partial sequence is generated automatically before the search by using empirically derived knowledge of specific amino acid contributions to fragment ion intensities in the spectrum. The program compares the sequences derived from the database search to the tandem mass spectrum using a dot product. GutenTag uses a fragmentation model for peptides that is based on the empirical observations and works well for doubly charged peptides derived from trypsin digestion. Sequence tagging approaches are useful for the identification of peptides with unknown modifications, amino acid sequence variations or sequence errors. Manual sequence interpretation can also lead to answers from spectra with unusual fragmentation that does not fit an algorithm's fragmentation model.

Stochastic models for database searching

Stochastic methods are based on probability estimates for peptide fragmentation and the subsequent generation of tandem mass spectra. In SCOPE[43], one of the early algorithms in this category, the MS/MS spectrum generation is modeled by a two-step stochastic process: fragmentation and measurement. The first step, fragmentation, enumerates all the possible fragmentation patterns of a peptide, and it determines the empirical probabilities associated with the pattern. The second step, measurement, generates tandem mass spectra from the fragments obtained in the first step, according to the distribution of the instrument measurement error.

The first step, the fragmentation probability estimation, allows the incorporation of physical and chemical properties of a peptide in the scoring process. A trivial fragmentation probability model would be, for example, to consider that the *b*- and *y*-ions of a peptide have a 100% chance of appearing in the MS/MS spectrum, the *a*-ions have a 50% chance and all the other possible fragment ions have a zero chance. Of course, the actual fragmentation process is more complex, and a more representative model can be derived from the analysis of large databases of manually curated spectra or by experienced mass spectrometrists.

Assuming that a large database of manually curated spectra is available, it is straightforward to generate a list of the observed fragment ions for each peptide. From here, the probability of appearance of each cleavage event in the tandem mass spectrum can be estimated by counting the number of times it is observed. In addition, using data mining techniques on this fragment data set, it is also possible to find those properties of a peptide that lead to significantly more cleavage events than would otherwise be expected.

Once the fragmentation probabilities have been estimated using one of the methods above, or a combination of the two, the second step of the algorithm is to compute the probability that a fragmentation pattern results in a tandem mass spectrum measurement. In

---

BOX 3 STRATEGY FOR LARGE SCALE DATA ANALYSIS

Large-scale proteomic experiments can present big challenges for data analysis. Most database searching algorithms can confidently identify amino acid sequences from tandem mass spectra showing good fragmentation and signal-to-noise ratio. Spectra of poorer quality or those containing aberrant fragmentation processes present the greatest challenge, as often the spectra are of peptides from low-abundance proteins. What are the strategies that can be used to mine a data set most thoroughly?

To limit the number of spectra and ensure an enriched set of unique spectra, poor-quality spectra could be removed and duplicates identified and eliminated[39,60,61]. Spectra should be searched with at least two algorithms to take advantage of the different selectivities of algorithms (for example, SEQUEST and Mascot). Unassigned spectra should be searched for modified amino acids. Any remaining spectra can be analyzed by automated or manual sequence tagging. Lastly, automated or manual *de novo* analysis can be applied to the remaining unassigned spectra.

To assist in the assignment of protein identifications, several algorithms have been developed to assess database searching results. Some programs simply filter the data and others assign a statistical confidence[24,55,62,63]. These programs are essential when dealing with large data sets.

BOX 4 FALSE POSITIVES IN DATABASE SEARCHING

False positives are a perpetual concern in database searching. They can arise for several reasons. Data-dependent algorithms for large-scale acquisition of tandem mass spectra do not discriminate between peptide ions and other types of ions that may be present. Thus, search algorithms are often confronted with a collection of spectra that could be single peptide ions, chemical noise, nonpeptide molecules or mixtures of coeluting isobaric peptides, which are then matched to amino acid sequences. Good data preprocessing or a search of a library of contaminants can help remove nonpeptide spectra prior to a search.

Peptides are often present at a wide range of concentrations in a sample, and peptides present at the limit of detection can produce poor quality fragmentation. The issue of sensitivity is more difficult to correct as it is heavily dependent on the limit of detection of a mass spectrometer. The effects can range from incomplete dissociation to poor ion statistics for fragment ions, making them indistinguishable from noise. In these cases incomplete fragmentation patterns or poor signal-to-noise ratios may lead to a solution that is not unique or correct.

The chemistry of peptide fragmentation is also not completely understood, and thus, fragmentation models used in database searching may not accommodate aberrant fragmentation processes and result in false positives. Several statistical studies of peptide fragmentation have been performed to better understand the contributions of specific amino acids to fragmentation processes. In time, improved models will account for more of the aberrant fragmentation processes.

Sequence conservation can lead to confusing results. If the same peptide sequence exists in multiple proteins, all of the proteins will be identified. Without additional peptide data it would be impossible to determine which protein produced the peptide that generated the tandem mass spectrum. Identifying this situation is straightforward, as most algorithms track all proteins that a spectrum matches.

A final possibility, and perhaps of more concern, are amino acid sequences that do not produce a unique fragmentation pattern but share enough of the same fragment ions to be indistinguishable from one another. In these cases a unique amino acid sequence can not be determined directly from the fragmentation pattern and other means are required to determine the absolute identity of the peptide. In particular, small peptides, less than eight amino acids in length, may not produce a fragmentation pattern that achieves a unique result.

---

other words, the probability of observing a collection of spectral peaks given a particular peptide fragmentation pattern is computed. The stochastic approach to this problem is to model the distribution of the measured m/z ratios, either as normal or uniform distributions centered at the m/z ratio for the predicted fragment.

Formally, the two-step process used by SCOPE can be described as follows: for a peptide p, a fragmentation pattern F and a tandem mass spectrum S, the first step of the algorithm estimates $Pr(F|p)$, the probability of obtaining the fragmentation pattern F from the collision-induced dissociation of peptide p. The second step of the algorithm determines $\psi(S|F,p)$, the probability of fragmentation pattern F to generate spectrum S. Finally, the probability of obtaining the spectrum S from peptide p is computed combining the two steps:

$$\psi(S,p) = \sum_{F \subseteq \Im(p)} \psi(S|F,p) Pr(F|p)$$

where $\Im(p)$ is the fragment space, which contains all of the fragmentation patterns theoretically generated from peptide p. SCOPE implements a dynamic programming algorithm to efficiently compute the above formula and reports the peptide p that maximizes the score, along with its corresponding P-value.

Stochastic models to determine the best fit between a tandem mass spectrum and a sequence have been used in other programs. The program OLAV uses an approach similar to the one used in SCOPE, except that some simplifications are made to avoid over-expansion of terms[51]. OLAV uses the maximum likelihood ratio as the identification score. The likelihood ratio is the ratio of probabilities from two alternative hypotheses, that peptide identification is valid and the alternative that the identification is random. A

program proprietary to Waters, MassLynx, uses Markovian chains to compute the statistical significance of a match, but a detailed description of the model used by the program is not available[58]. Most stochastic models require the use of training sets to determine the coefficients used to model features of the tandem mass spectrum. The magnitude of the coefficients may vary with the type of mass spectrometer used or other performance characteristics of the instrument, such as calibration. Therefore, the performance of these methods may be expected to depend on experimental settings, instruments used and limitations of the training set.

Statistical and probability models for database searching

This group of methods uses models based on empirically generated fragment ion probabilities[45,48,51]. In these methods no *a priori* determined probabilities are used. They generate a model that relates the sequences to a spectrum and determine the peptide identification score from this model. Thus, in the simplest models the frequencies of matches of *b*- and *y*-ions are determined and used to calculate a probability of sequence identification determined by the product of probabilities of its fragment matches. Several variations of this approach have been implemented in database searching algorithms[43,45,48,51]. Mascot[41] uses a model analogous to the one previously developed for identifying proteins from their peptide mass fingerprint[3]. Mascot may also use some empirical observations about fragment intensities and ion series continuity. The actual description of the model is not available in peer-reviewed literature and therefore we are not able to describe this algorithm in detail, even though it is one of the most widely used database search programs.

Recently, a group of database search algorithms have been implemented that use collective properties of database sequences to calculate the probability that a sequence match is a random event.

Thus, we have proposed to divide all database fragment ions into two groups: matches and misses[46]. Then, we assume that a hypergeometric probability models the frequencies of database peptides based on the number of matches. According to this model a probability that a peptide match is a random event is predicted from the hypergeometric probability of choosing K_1 matches (number of matches of a peptide) in N_1 trials (the number of fragment ions of the peptide) from a pool of fragments consisting of N fragments (number of all database fragments) K of which are matches (number of matches of all fragment ions to a spectrum). The hypergeometric probability of this event is:

$$P_{K,N}(K_1, N_1) = \frac{C_K^{K_1} \times C_{N-K}^{N_1 - K_1}}{C_N^{N_1}}$$

The probability of a peptide being a random match to the tandem mass spectrum is defined in the space that comprises all peptides whose mass match the mass of the precursor peptide. The significance of a peptide match is determined as a type I error of the null hypothesis—all fragment matches are random. OMSSA, a recently developed database search algorithm, uses a similar approach, where the peptide matches are modeled after the Poisson distribution[53]. Database search algorithms based on the number of matches trend to spectral quality owing to the fact that a match to a background peak and a match to a sequence ion are not distinguishable. Statistical models produce a statistical confidence for a match between the spectrum and database sequences. This confidence is based on the frequency of fragment ions in the database itself, and the probability a spectrum is a random match rather then the closeness of fit to a fragment model.

Conclusions

Automated analysis of tandem mass spectra is a critical process for new analytical strategies such as 'shotgun proteomics'. As tandem mass spectrometers have improved, the acquisition of hundreds of thousands of spectra has become not uncommon, and thus, accurate approaches to identify and validate sequence matches will make this method all the more powerful. Although a variety of algorithms have been demonstrated to provide accurate matches between tandem mass spectra and sequences, all suffer from an inability to provide verifiable matches to poor-quality spectra. Reliable and sensitive methods to assess spectral quality and assign quality indices to spectra will be critical for decreasing computational load and lowering false-positive rates (**Box 4**). Most algorithms are very accurate for peptides that follow general rules of fragmentation, but a subset of amino acid sequences and more highly charged peptide ions deviate from these rules; thus, a better understanding of relationships between peptide sequences and fragment ion intensity will assist in designing better models for matching spectra to sequences. Additional studies to better understand the strengths and weaknesses of the various algorithms will help to design algorithms with better sensitivity and selectivity.

1. Henzel, W.J. *et al.* Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90**, 5011–5015 (1993).
2. Yates, J.R., Speicher, S., Griffin, P.R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214**, 397–408 (1993).
3. Papin, D.J., Hojrup, P. & Bleasby, A.J. Rapid identification of proteins using peptide mass fingerprinting. Curr. Biol. 3, 327–332 (1994).
4. James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195**, 58–64 (1993).
5. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338–345 (1993).
6. Eng, J.K., McCormack, A.L. & Yates, J.R. III. 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).
7. Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
8. McCormack, A.L., Eng, J.K. & Yates, I. J. R. Peptide sequence analysis on quadrupole mass spectrometers. in *Methods: A Companion to Methods in Enzymology* **6**, 274–283 (1994).
9. McCormack, A.L., Eng, J.K., DeRoos, P.C., Rudensky, A.Y. & Yates, I. J. R. in *Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry* Vol. 619 (ed. Snyder, A.P.) 207–225 (American Chemical Society, Washington, D.C., 1995).
10. McCormack, A.L. *et al.* Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776 (1997).
11. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
12. Washburn, M.P., Wolters, D. & Yates, J.R. III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
13. Skop, A.R., Liu, H., Yates, J. III, Meyer, B.J. & Heald, R. Dissection of the mammalian midbody proteome reveals conserved cytokinesis mechanisms. *Science* **305**, 61–66 (2004).
14. Schirmer, E.C., Florens, L., Guan, T., Yates, J.R. III & Gerace, L. Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* **301**, 1380–1382 (2003).
15. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
16. Cheeseman, I.M. *et al.* Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p. *Cell* **111**, 163–172 (2002).
17. Sickmann, A. *et al.* The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. USA* **100**, 13207–13212 (2003).
18. 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).
19. Vihinen, M. Bioinformatics in proteomics. *Biomol. Eng.* **18**, 241–248 (2001).
20. Fenyo, D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11**, 391–395 (2000).
21. Fenyo, D. & Beavis, R.C. Informatics and data management in proteomics. *Trends Biotechnol.* **20**, S35–S38 (2002).
22. Yates, J.R. Database searching using mass spectrometry data. *Electrophoresis* **19**, 893–900 (1998).
23. Yates, J.R. III, McCormack, A.L. & Eng, J. Mining genomes with MS. *Anal. Chem.* **68**, 534A–540A (1996).
24. Nesvizhskii, A.I. & Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* **9**, 173–181 (2004).
25. Hunt, D.F., Yates, J.R. III, Shabanowitz, J., Winston, S. & Hauer, C.R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **83**, 6233–6237 (1986).
26. Papayannopoulos, I.A. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom. Rev.* **14**, 49–73 (1995).
27. Stults, J.T. & Watson, J.T. Identification of a new type of fragment ion in the collisional activation spectra of peptides allows leucine/isoleucine differentiation. *Biomed. Environ. Mass Spectrom.* **14**, 583–586 (1987).
28. Tabb, D.L. *et al.* Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75**, 1155–1163 (2003).
29. Schutz, F., Kapp, E.A., Simpson, R.J. & Speed, T.P. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc.*

Trans. **31**, 1479–1483 (2003).

30. Wysocki, V.H., Tsaprailis, G., Smith, L.L. & Breci, L.A. Mobile and localized protons: a framework for understanding peptide dissociation. _J. Mass Spectrom._ **35**, 1399–1406 (2000).

31. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. _Anal. Chem._ **76**, 3908–3922 (2004).

32. Mann, M., Meng, C.K. & Fenn, J.B. Interpreting mass spectra of multiply charged ions. _Anal. Chem._ **61**, 1702–1708 (1989).

33. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. & Pevzner, P.A. De novo peptide sequencing via tandem mass spectrometry. _J. Comput. Biol._ **6**, 327–342 (1999).

34. Sadygov, R.G. _et al._ Code developments to improve the efficiency of automated MS/MS spectra interpretation. _J. Proteome Res._ **1**, 211–215 (2002).

35. Colinge, J., Magnin, J., Dessingy, T., Giron, M. & Masselot, A. Improved peptide charge state assignment. _Proteomics_ **3**, 1434–1440 (2003).

36. Jonscher, K.R., Yates, I. & John, R. The quadrupole ion trap mass spectrometer—a small solution to a big challenge. _Anal. Biochem._ **244**, 1–15 (1997).

37. 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).

38. Tabb, D. Eng, JK, Yates, J.R. III in _Proteome Research: Mass Spectrometry_, Vol. 1 (ed. James, P.) 125–142 (Springer, New York, 2001).

39. Bern, M., Goldberg, D., McDonald, W.H. & Yates, J.R. III. Automatic quality assessment of peptide tandem mass spectra. _Bioinformatics_ **20** (Suppl. 1), I49–I54 (2004).

40. Fenyo, D., Qin, J. & Chait, B.T. Protein identification using mass spectrometric information. _Electrophoresis_ **19**, 998–1005 (1998).

41. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. _Electrophoresis_ **20**, 3551–3567 (1999).

42. 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).

43. Bafna, V. & Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. _Bioinformatics_ **17** (Suppl. 1), S13–S21 (2001).

44. Zhang, N., Aebersold, R. & Schwikowski, B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. _Proteomics_ **2**, 1406–1412 (2002).

45. Havilio, M., Haddad, Y. & Smilansky, Z. Intensity-based statistical scorer for tandem mass spectrometry. _Anal. Chem._ **75**, 435–444 (2003).

46. Sadygov, R. & Yates, J.R. I. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. _Anal. Chem._ **75**, 3792–3798 (2003).

47. Tabb, D.L., Saraf, A. & Yates, J.R. III. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. _Anal. Chem._ **75**, 6415–6421 (2003).

48. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P. & Gygi, S.P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. _Nat. Biotechnol._ **22**, 214–219 (2004).

49. Hansen, B.T., Jones, J.A., Mason, D.E. & Liebler, D.C. SALSA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. _Anal. Chem._ **73**, 1676–1683 (2001).

50. 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).

51. Colinge, J., Masselot, A., Giron, M., Dessingy, T. & Magnin, J. OLAV: towards high-throughput tandem mass spectrometry data identification. _Proteomics_ **3**, 1454–1463 (2003).

52. Field, H.I., Fenyo, D. & Beavis, R.C. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. _Proteomics_ **2**, 36–47 (2002).

53. Geer, L. in American Society for Mass Spectrometry (Nashville, Tennessee, USA, 2004).

54. MacCoss, M.J., Wu, C.C. & Yates, J.R. III. Probability-based validation of protein identifications using a modified SEQUEST algorithm. _Anal. Chem._ **74**, 5593–5599 (2002).

55. Sadygov, R.G., Liu, H. & 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).

56. Griffin, P.R. _et al._ Direct database searching with MALDI-PSD spectra of peptides. _Rapid Commun. Mass Spectrom._ **9**, 1546–1551 (1995).

57. Yates, J.R., Eng, J.K., Klausner, C. & Burlingame, A.L. Searching databases by using high energy CID spectra of peptides. _J. Am. Soc. Mass Spectrom._ **7**, 1089–1096 (1996).

58. Skilling, J. in EPTO, Vol. EP1047107 (Micromass, Europe; 1999). [AU: Please (1) give title of article (2) spell out 'EPTO'—is this a book or a journal?, and (3) if a book, please list editor(s), if any, and city and publisher]

59. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. _Biomed. Mass Spectrom._ **11**, 601 (1984).

60. Tabb, D.L., MacCoss, M.J., Wu, C.C., Anderson, S.D. & Yates, J.R. III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. _Anal. Chem._ **75**, 2470–2477 (2003).

61. Scherl, A. _et al._ Nonredundant mass spectrometry: a strategy to integrate mass spectrometry acquisition and analysis. _Proteomics_ **4**, 917–927 (2004).

62. Tabb, D.L., McDonald, H.W. & Yates, J.R. III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. _J. Proteome Res._ **1**, 21–36 (2002).

63. Kislinger, T. _et al._ PRISM, a generic large scale proteomic investigation strategy for mammals. _Mol. Cell. Proteomics_ **2**, 96–106 (2003).