ELSEVIER



# Software for computational peptide identification from MS–MS data

**Changjiang Xu and Bin Ma**

Department of Computer Science, University of Western Ontario, London, Ontario, N6A 5B7, Canada

**Protein identification in biological samples is an important task in drug discovery research. Protein identification is nowadays regularly performed by tandem mass spectrometry (MS–MS). Because of the difficulty of measuring intact proteins using MS–MS, typically a protein is enzymically digested into peptides and the MS–MS spectrum of each peptide is measured. Computational methods are then invoked to identify the peptides, which are later combined together to identify the protein. The most recognized peptide identification software packages can be classified into four categories: database searching, *de novo* sequencing, sequence tagging and consensus of multiple engines.**

The identification and quantification of proteins existing in a tissue are frequently key steps in the design of many proteomics and drug design investigations. Owing to the difficulty associated with the identification of intact proteins, they are often digested into short peptides and the individual peptides identified separately. Tandem mass spectrometry (MS–MS) is perhaps the most powerful analytical tool for protein and peptide identification [1,2]. The accuracy and speed of peptide identification are some of the key features that set MS–MS apart from the other methodologies used to analyze protein mixtures. As illustrated in Figure 1, peptide identification is a key computational step in identifying the proteins from MS–MS data. In fact, once the peptides are correctly identified, the later step of grouping the peptides and identifying the proteins becomes much simpler.

The principle of peptide identification using MS–MS is simple. A peptide is ionized and the peptide bonds are fragmented in an MS–MS spectrometer. Each type of resulting fragment ion will form a peak in the spectrum at the corresponding mass to charge ratio (m/z) of the ions. If a fragment ion has one more amino acid than another, the m/z difference between the two corresponding peaks will be equal to the mass of the amino acid divided by the charge state. A good quality spectrum might consist of a ladder of peaks of the y-ions (the fragment ions containing the carboxyl terminus) and a ladder of peaks of the b-ions (the fragment ions containing the amino terminus). Consequently, the peptide sequence can be derived by the mass differences of adjacent peaks in each of the two ladders. However, in practice, many factors complicate the problem. These include contamination of the sample, imperfect fragmentation, simultaneous fragmentation of two different peptides, post-translational modification (PTM) and low signal-to-noise ratio [3–5]. Consequently, in practice, many y-ion and b-ion peaks might be absent from, and many other types of peaks might unexpectedly appear in, the spectrum. These can make MS–MS peptide identification significantly harder than it would appear to be.

Over the past decade, numerous software programs have been developed for MS–MS peptide identification. These can be categorized into four classes: database searching, *de novo* peptide sequencing, peptide sequence tagging and consensus of multiple search engines. Given an MS–MS spectrum, database searching finds the best matching peptide from a protein sequence database; *de novo* sequencing computes a peptide directly from the spectrum; sequence tagging combines the two approaches by first conducting *de novo* sequencing to obtain a partial sequence (sequence tags), and then searches the sequence database using the sequence tags; consensus combines several different programs to increase the confidence and coverage. Software programs using these approaches are introduced below.

## Database searching

When the peptide of interest is known to be in a protein database, database searching is the most widely used approach for identification. In database searching software, the proteins in the database

*Corresponding author:* Ma, B. (bma@csd.uwo.ca)

1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2006.05.011

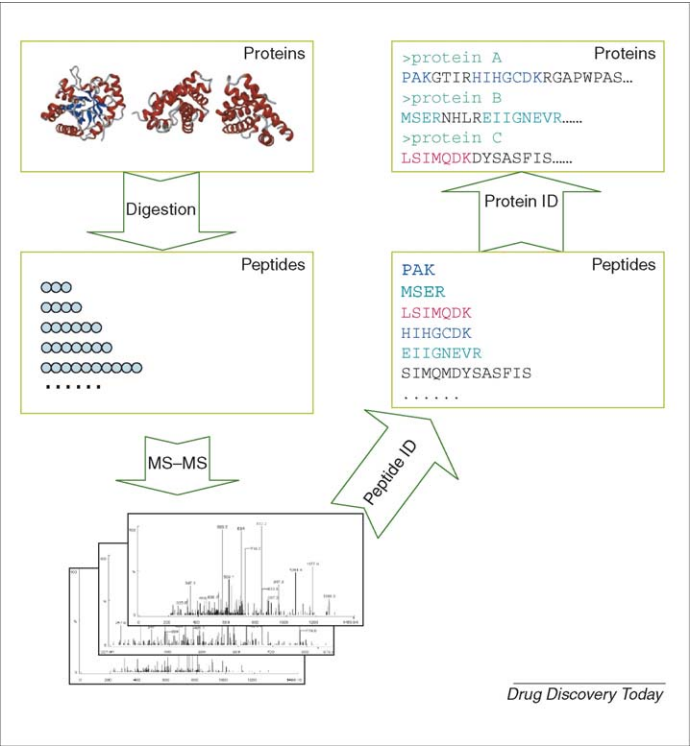


**FIGURE 1**

**The role of peptide identification in MS–MS protein analysis.** The downstream procedure illustrates the MS–MS experiments: purified proteins are digested into peptides and the MS–MS spectrum of each peptide is measured. The upstream procedure illustrates the computational steps in the data analyses: each spectrum is used to identify a peptide and then the peptides are grouped to identify proteins.

are digested virtually into peptides and each resulting peptide is compared with the input spectrum. Different software uses different criteria to determine the likelihood that the identified peptide is actually that seen in the spectrum. Ultimately, the most likely peptide is output as the result. The two common criteria are: (i) the peptide mass and (ii) the number and intensity of the peaks matched by the theoretically-computed m/z values of the fragment ions.

SEQUEST [6] is the earliest widely used software employing the database-searching approach. It uses a cross-correlation scoring function to evaluate matching between the spectrum and a database peptide sequence. The theoretical spectrum of the peptide sequence is computed using a simple model. The spectrum is then displaced by adding a displacement value $\tau$ to the m/z value of each peak. The correlation between the displaced theoretical spectrum and the experimental spectrum for each $-75 < \tau < 75$ is calculated and denoted by $f(\tau)$. The final score attributed to each peptide is equal to $f(0)$ minus the mean of $f(\tau)$. The difference between the scores of the first- and second-ranked peptide is a good discriminator between the correct identifications and false positives [6].

MASCOT [7] is currently the most widely used program for peptide identification. MASCOT uses the MOWSE scoring algorithm [8] to evaluate the match between a peptide and the input spectrum. The matches between the fragment ions of a database peptide and the peaks of the spectrum are regarded as random events. For each peptide, the probability that its matches occur randomly is calculated. The calculated probability should be exceedingly small for the correct peptide because many peaks will be matched. For better convenience, the probability $P$ is converted to a score of $-10 \times \log_{10}(P)$ before reporting. Because a protein database is not random, the MOWSE score only indicates the significance of the match and does not necessarily guarantee a correct match. As pointed out by the software manual, it is necessary to compare the score of the first-rank peptide with the other peptides to get a better idea on the correctness of the match.

PEAKS [3,9] was initially known as a *de novo* sequencing program. In its later versions, database searching was included. The software first conducts *de novo* sequencing to get a sequence for each spectrum. These sequences are then used to select many potential proteins from the protein database by sequence similarity. Finally, every spectrum is compared with every peptide of the potential proteins, using the same scoring function as used in the *de novo* sequencing. The score of the top-ranking database peptide is then compared with the other database peptides, as well as the *de novo* sequencing peptide, to compute a confidence score. As will be explained in the next section, *de novo* sequencing is equivalent to finding peptides in a 'universal' database that includes all linear amino acid combinations. Therefore, a scoring function that has enough discrimination power in *de novo* sequencing should work even better in a smaller database. In addition, comparing the top-ranking peptides in the real database and the 'universal' database is an effective way to remove false positives.

SEQUEST, MASCOT and PEAKS are all commercially available software. Free software programs have also been developed. Popular ones include Tandem [10] and OMSSA [11]. Tandem is a free open-source C++ program for rapid database searching. It breaks the search into two steps. A survey step makes some stringent assumptions about the peptides and rapidly identifies a set of protein sequences that are possible candidates. The candidates are then refined with a more time-consuming but more accurate scoring function that includes any evidence of incomplete enzymatic hydrolysis, nonspecific hydrolysis and chemical modifications of amino acid residues. Because the survey step restricts the later refinement step to a small set of proteins, this two-step strategy significantly speeds up the process. This two-step approach might have been used less explicitly in other database searching software. For example, the aforementioned PEAKS software also uses this two-step strategy but it identifies protein candidates by using sequence similarity. OMSSA [11] uses an explicit mathematical model for the matching probabilities. The E-value of the matching between a peptide and the input spectrum is calculated using the model, and peptides are ranked using the E-value. OMSSA is also an open source C++ program. Some programming techniques, such as memory mapped file, are used to make the program more efficient.

Other similar programs include MS-Tag, SCOPE, ProbID, OLAV, PFind, PepHMM, DBDigger and ProbIDtree [4,12–18]. These programs differ mainly by the scoring functions they use. PFind uses the correlative information for improving peptide identification accuracy. The kernel trick, rooted in the statistical learning theory, is exploited for scoring function. PepHMM combines information on machine accuracy, mass peak intensity and correlation among ions into a Hidden Markov Model (HMM) to calculate statistical significance of the HMM scores. DBDigger determines which

spectra can be compared with each candidate sequence, enabling the software to generate candidate sequences only once for each high-performance liquid chromatography separation, rather than for each spectrum, by reorganizing the database search process. ProbIDtree was designed to identify multiple peptides from a collision-induced dissociation spectrum generated by the concurrent fragmentation of multiple precursor ions by iterative database searching. Tentatively-matched peptides are organized in a tree structure from which their adjusted probability scores are calculated to determine the correct identifications.

## *De novo* peptide sequencing

A database-searching approach is only good for the identification of peptides present in a protein database. When an appropriate database is not available, *de novo* sequencing is the only way to identify the peptide. Besides the ability of identifying peptides without a database, *de novo* sequencing also has the advantages that: (i) *de novo* sequencing results can be used for homology-based database searches to identify peptide homologues and modifications, and (ii) *de novo* sequencing results can also be used to validate the database search results. Significant similarity between the database search result and the *de novo* sequencing result could be taken as evidence that the database-derived sequence is correct [5].

*De novo* sequencing can be regarded as identification of the peptide from the 'universal' peptide database that includes all linear amino acid combinations. Clearly, this is a more difficult problem than that encountered with database searching approaches. Although the scoring function is the most important factor in database searching software, *de novo* sequencing software also needs to incorporate an algorithm that can efficiently compute the optimal peptide under the scoring function. Searching every peptide in the 'universal' database is intractable with today's computers. The situation is further complicated because different amino acid combinations might have identical or nearly identical masses, and cleavages do not occur at every peptide bond, in which case the MS–MS spectrum will exhibit an incomplete series of b- and y-ions. With database searching approaches, the peptide might still be identified by the other b- and y-ions, whereas in *de novo* sequencing, the algorithm will have to examine the other ion types and the low-intensity peaks to figure out the missing information.

Despite the difficulties, many software programs have been developed. The speed and accuracy of *de novo* sequencing have also been improved significantly. Some of today's *de novo* sequencing programs, such as PEAKS [3,9] and PepNovo [19], can run at a speed of less than one second per spectrum on a personal computer. Lutefisk [20,21] was one of the earliest developed programs. PEAKS [3,9] has recently attracted attention because of its unique approach and excellent speed and accuracy. Other programs include Sherenga [22], SeqMS [23], Compute-Q [24], PepNovo [19] and NovoHMM [25]. Some of the earlier algorithms have previously been reviewed [26]. Most of these programs, except for PEAKS and NovoHMM, employ the spectrum graph approach for generating sequence candidates. PEAKS works on the spectrum directly instead of converting the spectrum to a graph. NovoHMM uses HMM. These programs are also differentiated by the scoring functions that evaluate the generated sequences.

In computer science, a graph is a very useful abstract data representation consisting of vertices and edges. The spectrum graph approach of *de novo* sequencing converts a spectrum into a graph, whereby each vertex corresponds to a possible ion related to a peak. Each edge connects two vertices whose corresponding ions have a mass difference approximately equal to the mass of an amino acid. Therefore, if the y-ion or b-ion ladders of the peptide appear in the spectrum, then there is a path (sequence of edges) that connects the amino and carboxyl termini. The spectrum graph approach computes the optimal peptide by computing the optimal path connecting the two termini. Earlier software programs using a spectrum graph approach used shortest-path [23] or heuristic algorithms [20,21] to compute the optimal path. Dancik *et al.* first claimed that there was an efficient algorithm to compute the 'antisymmetric longest path' in a spectrum graph [22]. The first efficient algorithm for the computation was published later by Chen *et al.* [24]. Many papers were published thereafter examining approaches to optimize the conversion from spectrum to graph to handle post-translational modifications and missing ions in the ladders.

Software programs employing the spectrum graph approach are summarized here. SeqMS [23] employs a single-source shortest-path algorithm to compute the path from the amino terminus to the carboxyl terminus. Lutefisk [20,21] traces sequences starting from the amino terminus until a sequence's mass matches the peptide molecular mass. Sherenga [22] transforms the experimental spectrum into a spectrum graph using ion types learned automatically by training. The peptide sequencing problem is then represented as the longest path problem in a directed acyclic graph. Compute-Q [24] uses a dynamic programming algorithm to find the longest antisymmetric path in the spectrum graph, derived as in Sherenga. PepNovo [19] uses a probabilistic network to calculate the likelihood that a y- or b-ion match is true; the logarithm of the likelihood ratio between the observed match and a random match is then used as the score of the ion match. The total ion match score is used to evaluate the candidate peptides. Extensive training is required to determine the parameters used in the probabilistic network. The current version of PepNovo only provides the parameters for charge-2 ion-trap mass spectrometers.

Unlike the spectrum graph model, which converts the spectrum to a graph, PEAKS software [3] uses a unique approach that works directly on the spectrum. The algorithm first computes a y-ion matching score and a b-ion matching score at each mass value according to the peaks around it. If there are no peaks around a mass value, a penalty value is assigned. The algorithm then efficiently computes many amino acid sequences that maximize the total scores at the mass values of b-ions and y-ions [9]. These candidate sequences are further evaluated by a more accurate scoring function, which also considers other ion types such as immonium ions and internal-cleavage ions [3]. The problem of ion absence is addressed because the PEAKS model assigns a score (or penalty) for each mass value. The software also computes a 'positional confidence' for each amino acid in the final result by examining the consensus of the top-scoring peptides.

A new approach, using an HMM, was proposed in NovoHMM [25]. HMM is a standard model that has been extensively used in other areas of bioinformatics and computer science. Each HMM includes some hidden states, some observable states and the

conditional probabilities that model the relationships between these states. Efficient algorithms exist to infer the best-possible hidden states from the observed states. NovoHMM carefully models the peptide sequence as hidden states and the spectrum as the observable states. The standard algorithm can then be adopted to infer the peptide sequence from the spectrum. One potential advantage of the approach of NovoHMM is that the scoring function combines the 'transition probability', which in this case is the probability that a certain amino acid follows another certain amino acid. Therefore, the accuracy will be boosted when the peptides are similar to the ones used to train the HMM. However, if the training is only done on a limited number of proteins, as described by Fischer *et al*. [25], it might give an exaggerated performance on these proteins but poor performance on the other proteins in the database.

## Peptide sequence tagging

Sequence-tagging approaches find the sequence of a peptide by searching a database with partial sequence information inferred from the MS–MS spectrum. The partial sequences are referred to as sequence tags. An example of a sequence tag is [258.1]TLMEY-LE[114.0]PK. The numerical values in the brackets represent amino acid combinations, with total mass equal to the values; however, the exact sequence in the brackets cannot be determined owing to lack of information in the spectrum. A sequence-tagging program will first infer tags from the spectrum by itself or by a separate *de novo* sequencing program, and then use these tags to select peptides and proteins from a protein database. In practice, even if the studied protein is not in a protein database, the chances are that the homologues of the proteins are in the protein database. In these cases, a sequence-tagging program that handles homology mutations can use the partial sequences to identify the protein homologues. A more sophisticated sequence-tagging program should also take care of the possible *de novo* sequencing errors existing in the tags.

Sequence tagging was first proposed for the error-tolerant peptide identification [27]. GutenTag [28] is another software program employing the sequence-tagging approach. Mann and Wilm [27] and Tabb *et al*. [28] made no special efforts to handle possible *de novo* sequencing errors in the partial sequences of the tags and homology mutations in the sequence database. To account for the homology mutations, a commonly employed method is to combine a *de novo* sequencing program such as PEAKS and a protein homology search program. Three general homology search programs, FASTA [29], Shotgun [30] and BLAST [31,32] have been modified to three sequence tag-searching programs: FASTS [33], MS-Shotgun [34], and MS-BLAST [35]. The weakness of these modified homology search programs is the ignorance of the possible *de novo* sequencing errors. Newer programs, such as OpenSea [36], SPIDER [37] and DeNovoID [38], consider these errors appropriately in their algorithms. In particular, SPIDER uses a rigorous algorithm to match a *de novo* sequence with a database sequence, enabling both *de novo* sequencing errors and homology mutations to occur at the same site of a peptide.

## Consensus of multiple search engines

None of the peptide-sequencing programs is perfect. Results generated by any single program, inevitably, have false positives and false negatives. It is tedious for a user to have to examine each of the thousands of peptide sequences to exclude false positives; and false negatives cause low coverage and identification confidence. Because more and more peptide-sequencing software programs are now available, researchers have started to use multiple programs to run the same dataset. The results of multiple engines are then combined to get fewer false positives, better coverage and higher confidence. In principle, different search engines give independent interpretations to the same data. If two or more independent interpretations produce the same result, then it is very likely that the result is correct. This will, clearly, help to reduce the number of false positives. Moreover, if three or more search engines are used, the coverage will be increased because data missing by one search engine might be picked up by another.

Although the use of multiple search engines has been a common practice in some laboratories for quite some time, the software for performing this analysis automatically has only recently become available. This approach was first used by Resing *et al*. [39] to combine the search results of MASCOT and SEQUEST. Scaffold [40] is one of the first programs dedicated to combining the results of multiple search engines. It assigns confidence scores to the results of different search engines statistically and then computes the confidence score of the combined result. PEAKS software has also recently added this consensus functionality. It performs searches using multiple search engines and then combines results to generate a unique report [41]. The results reported by Resing *et al*. [39], Searle [40] and Rogers [41] have shown that multiple search engines can result in a dramatic improvement in peptide identification accuracy, coverage and confidence.

## Conclusion and discussion

The recent development of both the hardware and software in MS–MS spectrometry has enabled high-throughput peptide identification using MS–MS. Numerous peptide identification programs have been developed. The methods used in these programs can be categorized into four classes: database searching, *de novo* sequencing, peptide tagging and consensus. Database searching selects a peptide from a protein sequence database to best match the input spectrum; d*e novo* sequencing computes such a peptide from the spectrum without using a sequence database; peptide tagging uses *de novo* sequencing results to find the peptides or their homologues in the sequence database; the consensus approach combines the search results of multiple programs to attain fewer false positives, better coverage and higher confidence. It is noteworthy that some software might belong to multiple categories. In particular, the PEAKS software implements all of the four approaches.

Most software programs reviewed in this article can support the identification of peptides with PTM. However, these programs usually perform a restrictive search that takes into account only several types of PTM. More recently, a blind PTM search algorithm, MS-Alignment, has been developed [42]. Many of the reviewed programs can support the input of MS–MS data from multiple types of MS–MS instruments. Table 1 lists the availabilities and characteristics of some software packages reviewed in this article.

Different publications have often claimed different software to be superior to the others. Also, many software packages, especially the commercial ones, undergo constant improvement through version upgrades. For these reasons, this article did not attempt to

**TABLE 1**

**The availabilities and characteristics of the most recognized software packages for peptide identification[a,b]**

| Software name | Availability | Website | Software type | | Function | | | | Supported instrument[c] | | | | PTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Online | Desktop | Database searching | *De novo* | Peptide tagging | Consensus | Iontrap | QTof | FTMS | FT (ECD) | User-defined | Standard |
| PEAKS | Commercial, free online | http://www.bioinfor.com/peaksonline | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| MASCOT | Commercial, free online | http://www.matrixscience.com | Y | – | Y | Y | Y | – | Y | Y | Y | Y | – | Y |
| SEQUEST | Commercial | http://fields.scripps.edu/sequest | – | Y | Y | – | – | – | Y | Y | Y | Y | Y | Y |
| Tandem | Open source, free | http://www.thegpm.org/tandem | Y | Y | Y | – | – | – | Y | Y | Y | – | Y | Y |
| OMSSA | Open source, free | http://pubchem.ncbi.nlm.nih.gov/omssa | Y | Y | Y | – | – | – | Y | – | – | – | – | Y |
| Lutefisk | Open source, free | http://www.hairyfatguy.com/Lutefisk | – | Y | – | Y | – | – | Y | Y | – | – | Y | Y |
| PepNovo | Open source, free | http://peptide.ucsd.edu/pepnovo.py | Y | Y | – | Y | – | – | Y | – | – | – | Y | Y |
| MS-BLAST | Free online | http://dove.embl-heidelberg.de/Blast2/msblast.html | Y | – | – | – | Y | – | | | | | – | – |
| SPIDER | Free online | http://bif.csd.uwo.ca/spider | Y | Y | – | – | Y | – | | | | | – | Y |
| Scaffold | Commercial | http://www.proteomesoftware.com | – | Y | – | – | – | Y | Y | Y | – | – | | |

[a] An 'online' software type indicates that the software can run as a web server and be accessed through a web browser remotely. The 'PTM' (post-translational modification) column indicates whether the software can identify peptides with post-translational modifications.
[b] Y, available; –, not available; white space, not applicable.
[c] Software can normally process data from all three types of instruments (Iontrap, QTof and FTMS) as long as the data are converted to an appropriate file format. However, some software does not currently have parameters specially trained for certain instrument types, and therefore cannot fully utilize the data.

compare the performances of different software. However, all of the software packages listed in Table 1 are well recognized. Some comparative studies have also been carried out by Shadforth *et al.* [5] and Kapp *et al.* [43]. Recently, there have also been efforts, such as PeptideProphet [44], to validate peptide identification results using automated computational methods. The validation tools have been reviewed elsewhere [45].

In addition to the software reviewed here, mass spectrometer vendors often provide peptide identification software with their instruments. Usually, their software uses methods that have not been published. Therefore, these software programs (with the exception of SEQUEST) are not reviewed here. A common difficulty of MS–MS data analysis and sharing is that mass spectrometer vendors have their proprietary data formats. Public efforts have been made to standardize the data formats using XML. These include the mzXML format developed by the Sashimi project [46] and the mzData format proposed by the Human Proteome Organization.

## References


1 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
2 Snyder, A.P. (2000) *Interpreting Protein Mass Spectra: A Comprehensive Resource.* Oxford University Press
3 Ma, B. *et al.* (2003) PEAKS: powerful software for peptide *de novo* sequencing by MS/MS. *Rapid Commun. Mass Spectrom.* 17, 2337–2342
4 Zhang, N. *et al.* (2005) ProbIDtree: 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
5 Shadforth, I. *et al.* (2005) Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics* 5, 4082–4095
6 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
7 Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* 20, 3551–3567
8 Pappin, D.J.C. *et al.* (1996) Chemistry, mass spectrometry and peptide-mass databases: evolution of methods for the rapid identification and mapping of cellular proteins. *Mass Spectrom. Biol. Sci.* 135–150
9 Ma, B. *et al.* (2005) An effective algorithm for the peptide *de novo* sequencing from MS/MS spectrum. *J. Comput. Syst. Sci.* 70, 418–430
10 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
11 Geer, L.Y. *et al.* (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* 3, 958–964
12 Clauser, K.R. *et al.* (1996) Peptide fragment-ion tags from maldi/psd for error-tolerant searching of genomic databases. *44th ASMS Conf. Mass Spectrom. and Allied Topics*, Portland, OR, USA. May 12–16, pp. 365
13 Bafna, V. and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 17, S13–S21
14 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
15 Colinge, J. *et al.* (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* 3, 1454–1463
16 Fu, Y. *et al.* (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* 20, 1948–1954
17 Wan, Y. *et al.* (2005) PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *RECOMB* 342–356
18 Tabb, D.L. *et al.* (2005) DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal. Chem.* 77, 2464–2474

19 Frank, A. and Pevzner, P. (2005) Pepnovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 77, 964–973

20 Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11, 1067–1075

21 Taylor, J.A. and Johnson, R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 73, 2594–2604

22 Dancik, V. *et al.* (1999) *De novo* peptide sequencing via tandem mass-spectrometry. *J. Comput. Biol.* 6, 327–342

23 Fernandez-de Cossio, J. *et al.* (1995) A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput. Appl. Biosci.* 11, 427–434

24 Chen, T. *et al.* (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 8, 325–337

25 Fischer, B. *et al.* (2005) NovoHMM: a hidden Markov model for *de novo* peptide sequencing. *Anal. Chem.* 77, 7265–7273

26 Lu, B. *et al.* (2004) Algorithms for *de novo* peptide sequencing via tandem mass spectrometry. *Drug Discov. Today: BIOSILICO* 2, 85–90

27 Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399

28 Tabb, D. *et al.* (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* 75, 6415–6421

29 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448

30 Pegg, S.C. and Babbitt, P.C. (1999) Shotgun: getting more from sequence similarity searches. *Bioinformatics* 15, 729–740

31 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

32 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

33 Mackey, A.J. *et al.* (2002) Getting more for less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics* 1, 139–147

34 Huang, L. *et al.* (2001) Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* 276, 28327–28339

35 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

36 Searle, B.C. *et al.* (2004) High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. *Anal. Chem.* 76, 2220–2230

37 Han, Y. *et al.* (2005) SPIDER: software for protein identification from sequence tags with *de novo* sequencing error. *J. Bioinform. Comput. Biol.* 3, 697–716

38 Halligan, B.D. *et al.* (2005) DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from *de novo* peptide sequencing by mass spectroscopy. *Nucleic Acids Res.* 33, 376–381

39 Resing, K.A. *et al.* (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* 76, 3556–3568

40 Searle, B.C. *et al.* (2005) Scaffold: a program to probabilistically combine results from multiple MS/MS database search engine. Association of Biomolecular Resource Facilities, ABRF'05. Savannah, GA, USA, Abstract #P87-T

41 Rogers, I. (2005) Assessment of an amalgamative approach to protein identification. ASMS Conference on Mass Spectrometry. San Antonio, TX, USA, 5–9 June 2005. Abstract #WP379

42 Tsur, D. *et al.* (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* 23, 1562–1567

43 Kapp, E.A. *et al.* (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 5, 3475–3490

44 Keller, A. *et al.* (2002) Empirical statistical model to evaluate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.* 74, 5383–5392

45 Nesvizhskii, A.I. and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* 9, 173–181

46 Pedrioli, P.G.A. *et al.* (2004) A common open representation of mass spectrometry data and its application in a proteomics research environment. *Nat. Biotechnol.* 22, 1459–1466