

Algorithms for *de novo* peptide sequencing using tandem mass spectrometry

Bingwen Lu and Ting Chen

There is growing interest in the qualitative and quantitative analysis of proteins on a proteome-wide scale. Mass spectrometry plays an important role in the high-throughput study of proteomics. There have been two major advances in mass spectrometry technology: instrumental (including physicochemical) development, such as the development of chromatography, protein ionization and protein labeling technologies; and the development of computational algorithms for the analysis of mass spectra produced by mass spectrometers. In this review, we provide an overview of the development of computational algorithms for *de novo* peptide sequencing using tandem mass spectrometry.

Bingwen Lu
Ting Chen*

Molecular and Computational
Biology Program
Department of Biological
Sciences
University of Southern California
Los Angeles
CA 90089, USA

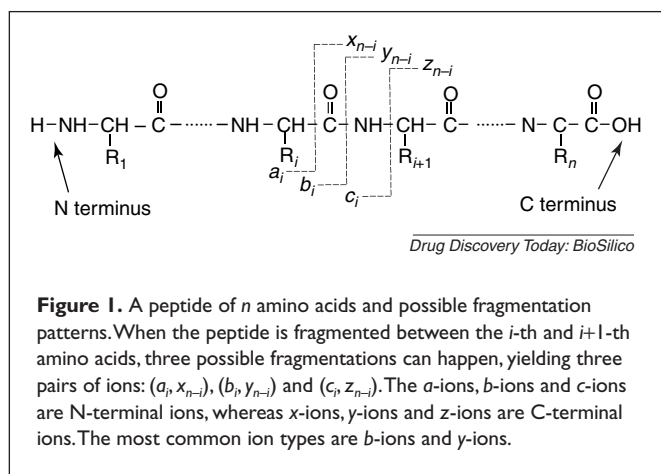
*e-mail: tingchen@hto.usc.edu

▼ Mass spectrometry is an indispensable tool in the era of modern proteomics research. Mass spectrometry has been applied successfully to the proteome-wide study of proteins (e.g. [1–6]). For example, Gavin *et al.* [3] used tandem affinity purification (TAP) and mass spectrometry to characterize multiprotein complexes in *Saccharomyces cerevisiae*. They processed 1739 genes and purified 589 protein assemblies. Their analysis of these assemblies revealed 232 distinct multiprotein complexes. They proposed cellular roles for 344 proteins, among which 231 had no previous functional annotation. In addition, Ho *et al.* [4] employed a method called high-throughput mass spectrometric protein complex identification (HMS-PCI) to systematically identify protein complexes in *S. cerevisiae*. Starting with 10% of the predicted proteins, they detected 3617 associated proteins, covering 25% of the yeast proteome. A third example is the analysis by Petricoin *et al.* [5] of the proteomic pattern in serum by mass spectrometry for ovarian cancer diagnosis. A training set of mass spectra generated from 50 unaffected women and 50 ovarian cancer patients was analyzed by an

artificial intelligence algorithm to discover a proteomic pattern that could completely discriminate cancer samples from normal samples. The identified pattern was then used to classify masked serum samples. The authors successfully identified all 50 ovarian cancer samples, whereas for the 66 cases of nonovarian cancer 63 were classified as noncancer. They computed the positive predictive value of this validation to be 94%. This outperformed a different ovarian cancer diagnostic measure called CA125. The CA125 blood test is used to measure the level of CA125, a cancer-related antibody and the most common tumor marker in ovarian cancer, and has a positive predictive value of 34%.

Protein sequencing and identification by tandem mass spectrometry

When sequencing and identifying proteins by tandem mass spectrometry, the protein or proteins are first digested by enzymes such as trypsin. The resulting peptides are then separated by liquid chromatography (LC) and subsequently analyzed by a mass spectrometer. The peptides are ionized and the mass-to-charge (m/z) ratios measured. For tandem mass spectrometry, ions within a given range of specific m/z ratios are selected and fragmented, and the m/z is measured. The resulting tandem mass spectra, comprising m/z ratios and corresponding intensities, are then used for the identification of the original protein or peptide. There are two principal ways of doing this kind of identification. One is to correlate the spectra with protein sequences or nucleic acid sequences by database searching [7–12]. The other is to derive the peptide sequence directly, without the help of any



sequence database. The latter method is called *de novo* peptide sequencing.

The *de novo* peptide sequencing problem and the scoring function

When presented with a tandem mass spectrum, the goal is to seek the true peptide whose fragmentation gave rise to the spectrum in question. Often this is reduced to finding a peptide that can best explain the spectrum according to a scoring function. This requires a discussion of the meaning of a scoring function. For a given real experimental spectrum, we can find candidate peptides. Once the candidate peptides are obtained, a hypothetical spectrum can be generated *in silico* for each of these, based on the assumption of some fragmentation patterns. Different fragmentation patterns will generate different ion types, as shown in Figure 1. The most common ion types are b -ions and y -ions. The scoring function then measures the similarity of the experimental spectrum to the hypothetical spectrum of each candidate peptide. The candidate peptide associated with the hypothetical spectrum that shows the most similarity to the experimental spectrum is then reported as the best candidate peptide. Sometimes the similarity score is associated with a p -value, which gives the probability that the score is achieved by random chance (e.g. see [10]). The design of a good scoring function is never an easy task and is an active research area [7–14].

A brief description of one of the scoring methods, a four-step process called the SEQUEST algorithm [7], is given to illustrate the abstract description of the scoring function. The algorithm begins (step 1) with tandem mass spectrometry data reduction. In this step, fractional m/z ratios are rounded to the nearest integers and then a 10 unit (u) window around the precursor ion (or parent ion) is removed to avoid matching to the unfragmented precursor ion. The 200 most abundant ions are selected for scoring purposes.

In step 2, the candidate peptide sequences are then identified by matching the precursor peptide masses to the masses of all possible peptides in a protein database. Peptides with ± 3 u or ± 1 u are selected as candidate peptides. One hypothetical spectrum is then generated for each candidate peptide and these hypothetical spectra are compared with the experimental spectrum to produce a preliminary ranked list of 500 best-fit sequences (step 3). This preliminary ranked list considers the number of matching ions within the mass tolerance of ± 1 u, the abundances of the matching ions, the continuity of an ion series, and the presence of immonium ions for the amino acids histidine, tyrosine, tryptophan, methionine and phenylalanine. Finally, in step 4, these 500 sequences are then subject to a cross-correlation analysis to generate the final ranked list.

Rationales for *de novo* sequencing

Usually, searching a sequence database is the first choice for peptide identification. However, *de novo* peptide sequencing comes into play in various situations. First, the protein of interest might not be present in the sequence database. This could be because the sequence database is incomplete, which is the situation for many model animals and plants. It could also be due to the fact that the protein of interest is a novel protein. Second, there are prediction errors in gene-finding programs. Thus, it might not be possible to find the true protein from the predicted protein database. Third, some scientists might want to study the proteome before the genome, in which case no sequence database would be available. Fourth, genes might undergo alternative splicing, which would result in novel proteins. The occurrence of single nucleotide polymorphisms (SNPs) in coding regions may also lead to different protein variants. Fifth, *de novo* sequencing can be helpful for studying amino acid mutations and protein modifications. Finally, when a database search generates ambiguous results, *de novo* sequencing can be used as a validation tool.

Early development of *de novo* sequencing algorithms

Over the years, various algorithms have been developed to address the *de novo* sequencing problem. One naive method [15,16] is to list all possible candidate peptides according to the mass of the parent ion of the tandem mass spectrum. This is sometimes called 'exhaustive listing'. All of the candidate peptides are then compared with the experimental spectrum to find out which one is the best match. One computational difficulty inherent to this approach is that there will most probably be a large number of possible candidates for typical parent masses that may

range from less than one thousand Daltons to several thousand Daltons [15,16]. For example, as described in [15], for a parent peptide of molecular weight 774, there will be 21 909 046 possible candidate peptides.

An alternative approach, sometimes called 'subsequencing', has also been tried on mass spectra data generated by various mass spectrometers [17–20]. In this approach, short sequences that represent only a portion of the whole sequence are tested against the experimental spectrum. Those subsequences that account for the observed ions are then extended one residue at a time until the whole sequence is tested. During subsequence extension, only those subsequences that significantly match the experimental spectrum are retained. One disadvantage of this approach is that some good candidate peptides might be discarded when some regions of a peptide are less represented by fragmentation ions. It is important to be aware that the fragmentation frequencies of a peptide are usually not evenly distributed over the whole peptide.

A third method employs graphical display of the data [21]. In this method, fragmentation ions that differ by the mass of one amino acid are represented by connected lines, thus allowing the visualization of ion series of the same type. Such an approach is not suitable for high-throughput environments, but this method can be quite helpful for manual *de novo* interpretation of tandem mass spectra.

A fourth approach uses graph theory [13,22–25]. This approach has proven to be quite successful and will be discussed in the next section.

Application of graph theory in mass spectrometry

The application of graph theory to *de novo* peptide sequencing was first proposed by Bartels [22]. The basic idea is to transform an experimental spectrum into a graph called a 'spectrum graph'. In the transformation, each peak in the experimental spectrum is represented as a vertex (or several vertices) in the spectrum graph and a directed edge is established between two vertices if the mass difference of the two vertices equals the mass of one or several amino acids. Various algorithms have been designed to find paths in the spectrum graph for which the corresponding peptides provide a good explanation of the experimental spectrum. Some algorithms of this type are introduced below.

The SeqMS algorithm

The SeqMS algorithm was originally called MSEQ [24]. First, a list of possible ion types with corresponding probabilities was assumed. The list was then used to transform the experimental spectrum into a spectrum graph. Each peak will correspond to a set of vertices in the spectrum graph, according to the list of ion types. A graph is then obtained by linking

all pairs of vertices that differ by the mass of an amino acid or the combination of several amino acids. Dijkstra's single-source, shortest-path algorithm was then employed to find the complete path from the N terminus to the C terminus. The SeqMS program is available at <http://www.protein.osaka-u.ac.jp/rcsfp/profiling/Seqms/SeqMS.html>

The Lutefisk algorithm

The Lutefisk algorithm was designed by Taylor and Johnson [25]. In this algorithm, the authors first reduced the experimental spectrum data to a list of significant fragment ions. They then determined the N- and C-terminal evidence lists, which reveal the possible N- and C-terminal ions, respectively. After the N- and C-terminal evidence lists were obtained, a 'sequence spectrum' was formed, in which the x ordinate consisted of the m/z values for the b -ions and the y ordinate consisted of the probability of cleavage at each site. The program then proceeded by tracing out sequences, starting from the N terminus, by finding b -ion values that differed from the N-terminal ion by the mass of one or several amino acids. After all the sequences had been obtained, a scoring procedure was used to rank the sequences. The Lutefisk program was once publicly available. The program was removed from the public domain after Immunex was acquired by Amgen.

The Sherenga algorithm

The Sherenga algorithm was developed by Dancik *et al.* [13]. Because the creation of a spectrum graph is based on the ion types, the authors designed a method to automatically learn ion types from a training set of experimental spectra of known sequences, without knowing *a priori* the fragmentation patterns. After the ion types were learned, they transformed the experimental spectrum into a spectrum graph using the following steps. First, the ion types were represented by $\Delta = \{\delta_1, \dots, \delta_k\}$, where k is the number of ion types and each δ_i represents the offset of the corresponding ion type. The experimental spectrum S was then transformed into a spectrum graph $G_\Delta(S)$ as follows. Each peak s of the experimental spectrum S generated k vertices $V(s) = \{s + \delta_1, \dots, s + \delta_k\}$. Two vertices, u and v , were then connected by a directed edge from u to v if $v - u$ equaled the mass of an amino acid. The peptide sequencing problem was then represented as the longest path problem in a directed acyclic graph. The authors also pointed out that the longest path might correspond to unrealistic solutions because it may use multiple vertices associated with the same experimental spectral peak. One solution is to find the longest antisymmetric path. They claimed that an efficient algorithm exists to find the longest antisymmetric path in the spectrum graph, but did not describe the algorithm.

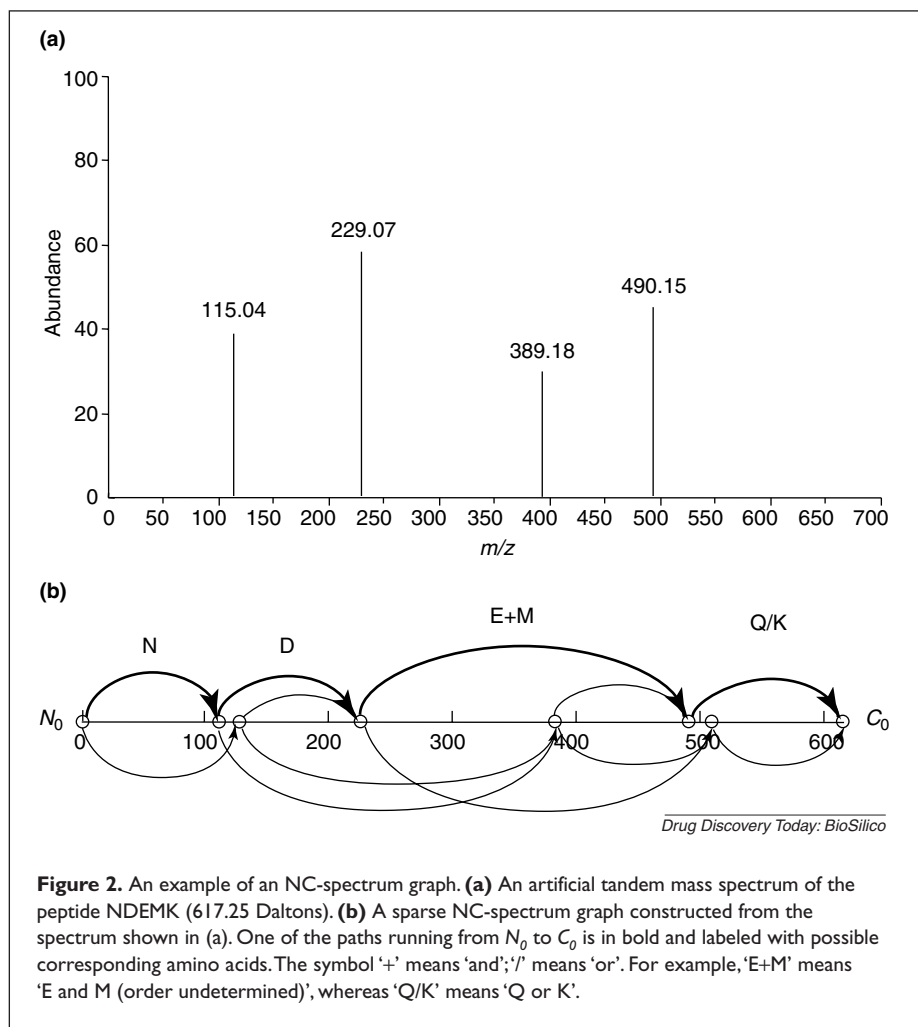


Figure 2. An example of an NC-spectrum graph. **(a)** An artificial tandem mass spectrum of the peptide NDEMK (617.25 Daltons). **(b)** A sparse NC-spectrum graph constructed from the spectrum shown in (a). One of the paths running from N_0 to C_0 is in bold and labeled with possible corresponding amino acids. The symbol '+' means 'and'; '/' means 'or'. For example, 'E+M' means 'E and M (order undetermined)', whereas 'Q/K' means 'Q or K'.

optimal solution to the original problem. Chen *et al.* [28], employing dynamic programming, provided a polynomial time algorithm for the *de novo* sequencing problem using a spectrum graph. First, the NC-spectrum graph G (NC denotes N-terminal and C-terminal) is constructed from the experimental spectrum, by assuming that each peak in the experimental spectrum can be a *b*-ion or a *y*-ion. Thus, each peak in the experimental spectrum will generate two vertices in the NC-spectrum graph G . All of the vertices are then placed on the real line at positions corresponding to the mass values of the vertices. If the mass difference between two vertices u and v equals the total mass of one or several amino acid residues, a directed edge is drawn between u and v , pointing from the low-mass vertex to the high-mass vertex. To give the reader a feel for how an NC-spectrum graph can be generated from a mass spectrum, an example is given in Figure 2. Next, the nodes of G are renamed in order from left to right as $(x_0, x_1, \dots, x_k, y_k, \dots, y_1, y_0)$, where every pair, x_i and y_i , $1 \leq i \leq k$, corre-

Dynamic programming and suboptimal concept

As pointed out by Dancik *et al.* [13], the problem of finding the longest path in a directed acyclic spectrum graph while avoiding multiple assignments to the same peak is NP-complete in the general case [26]. NP-complete problems are a class of hard problems such that, if any NP-complete problem can be solved in polynomial time, then every NP-complete problem has a polynomial time algorithm [27]. However, Dancik *et al.* [13] and Chen *et al.* [28] observed that there is a special structure for forbidden pairs of vertices (twins) in the spectrum graph. That is, the forbidden pairs are noninterleaving. Two forbidden pairs of vertices x_1, y_1 and x_2, y_2 are noninterleaving if the intervals x_1, y_1 and x_2, y_2 do not interleave. Chen *et al.* [28] then proposed a dynamic programming approach to find the longest anti-symmetric path in the spectrum graph.

Dynamic programming is a common technique for solving optimization problems [26]. In dynamic programming, an optimization problem is solved in a bottom-up fashion by combining the solutions to subproblems, which gives an

sponds to two mutually exclusive assumptions of the same mass peak. The dynamic programming algorithm then finds the path with the maximum path score from x_0 to y_0 that contains the edge (x_i, y_i) , $i \neq j$.

The dynamic programming algorithm will find the optimal solution. However, the optimal solution may not be the sequence that produces the experimental spectrum. Even database search programs sometimes report several sequences with similar scores because the scoring function can misinterpret the spectral data. Noise and unknown ions may also be interpreted as real ions by the programs. For these reasons, the suboptimal solutions are of great interest because they might give us the actual sequence that generated the spectrum. Lu and Chen [29] further explored this application of suboptimal solutions in the dynamic programming algorithm. In this suboptimal algorithm, an experimental spectrum with k peaks is transformed into a matrix spectrum graph $G = (V, E)$, where $|V| = O(k^2)$ and $|E| = O(k^3)$. A polynomial time suboptimal algorithm was then proposed to find all of the suboptimal solutions

(peptides) in $O(p|E|)$ time, where p is the number of solutions. The suboptimal algorithm has been implemented and is available at <http://hto-c.usc.edu:8000/msms/>

Following the work of Dancik *et al.* [13] and Chen *et al.* [28], two more research groups also proposed dynamic algorithms to solve the *de novo* peptide sequencing problem [30,31]. Bafna and Edwards [30] also considered the suboptimal solutions in their dynamic programming machinery. The work by Ma *et al.* [31] is basically a dynamic programming approach and thus a description is omitted here. However, a link to their implementation is provided at <http://www.bioinformaticssolutions.com/software/peaks/>

Besides the algorithms mentioned above, there are also some commercial software packages for *de novo* sequencing. For example, Thermo Finnigan has recently released the DeNovoX™ software package for automated *de novo* peptide sequencing. Micromass provides a *de novo* sequencing tool called MassSeq™. MassSeq™ uses a Bayesian approach to score randomly generated peptides to find the best possible match with the tandem mass spectrometry data.

Concluding remarks

Mass spectrometry has become an important tool for proteomics. Normally, searching against a sequence database is the first choice for protein identification. However, *de novo* sequencing can also be used in various situations. Over the years, numerous computer algorithms have been developed for *de novo* peptide sequencing and there are also manual methods for the interpretation of mass spectra. For example, one manual strategy for the interpretation of mass spectra generated from tryptic peptides was presented by Kinter and Sherman [32]. There are also novel methods that employ multistage tandem mass spectrometry (MS^n) for *de novo* peptide sequencing, whereby some ions of MS^2 are identified by MS^3 and even MS^4 [33,34].

At the same time, the *de novo* peptide sequencing problem using tandem mass spectrometry is still not solved in general. In other words, the information revealed by tandem mass spectrometry cannot be readily converted into a fully unambiguous peptide sequence.

Usually, a *de novo* sequencing program has been designed for given machine-dependent tandem mass spectra and thus is not universally applicable to spectra generated by other types of mass spectrometers. To our knowledge, none of the current *de novo* sequencing programs takes into account internal fragmentation of the parent ion.

Nevertheless, the development of computational algorithms, including *de novo* peptide sequencing methods, database search algorithms and other computational tools, together with mass spectrometric instrumental

developments, would substantially enhance our capabilities in biological studies.

Acknowledgements

We thank Dr. Chris Watson and the three reviewers for helpful comments on the manuscript.

References

- Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682
- Washburn, M.P. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 19, 242–247
- Gavin, A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
- Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
- Petricoin, E.F. *et al.* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572–577
- 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
- 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
- Mann, M. and Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399
- Clauser, K.R. *et al.* (1999) Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS. *Anal. Chem.* 71, 2871–2882
- Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567
- Bafna, V. and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 17 (suppl 1), S13–S21
- Field, H.I. *et al.* (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* 2, 36–47
- Dancik, V. *et al.* (1999) *De novo* peptide sequencing via tandem mass spectrometry: a graph-theoretical approach. *J. Comput. Biol.* 6, 327–342
- Havilio, M. *et al.* (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* 75, 435–444
- Sakurai, T. *et al.* (1984) PAAS 3, a computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Mass Spectrom.* 11, 396–399
- Hamm, C.W. *et al.* (1986) Peptide sequencing program. *Comput. Appl. Biosci.* 2, 115–118
- Biemann, K. *et al.* (1966) Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.* 88, 5598–5606
- Ishikawa, K. and Niwa, Y. (1986) Computer-aided peptide sequencing by fast-atom-bombardment mass-spectrometry. *Biomed. Environ. Mass Spectrom.* 13, 373–380
- Siegel, M.M. and Bauman, N. (1988) An efficient algorithm for sequencing peptides using fast atom bombardment mass-spectra data. *Biomed. Environ. Mass Spectrom.* 15, 333–343
- Johnson, R.S. and Biemann, K. (1989) Computer-program (seqpep) to aid in the interpretation of high-energy collision tandem mass-spectra of peptides. *Biomed. Environ. Mass Spectrom.* 18, 945–957
- Scoble, H.A. *et al.* (1987) A graphics display-oriented strategy for the amino-acid sequencing of peptides by tandem mass-spectrometry. *Fresenius Zeitschrift Fur Analytische Chemie* 327, 239–245

RESEARCH FOCUS

- 22 Bartels, C. (1990) Fast algorithm for peptide sequencing by mass spectrometry. *Biomed. Environ. Mass Spectrom.* 19, 363–368
- 23 Hines, W.M. *et al.* (1992) Pattern-based algorithm for peptide sequencing from tandem high-energy collision-induced dissociation mass-spectra. *J. Am. Soc. Mass Spectrom.* 3, 326–336
- 24 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
- 25 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
- 26 Cormen, T.H. *et al.* (2001) *Introduction to Algorithms*, The MIT Press
- 27 Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company
- 28 Chen, T. *et al.* (2001) A dynamic programming approach for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 8, 325–337
- 29 Lu, B. and Chen, T. (2003) A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 10, 1–12
- 30 Bafna, V. and Edwards, N. (2003) On *de novo* interpretation of tandem mass spectra for peptide identification. Annual Conference on Research in Computational Molecular Biology
- 31 Ma, B. *et al.* (2003) An effective algorithm for the peptide *de novo* sequencing from MS/MS spectrum. *CPM* 266-277
- 32 Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons
- 33 Carte, N. *et al.* (2001) *De novo* sequencing by nano-electrospray multiple-stage tandem mass spectrometry of an immune-induced peptide of *Drosophila melanogaster*. *Eur. J. Mass Spectrom.* 7, 399–408
- 34 Macht, M. *et al.* (2001) Mass spectrometric analysis of head-to-tail connected cyclic peptides. *Acta Biochim. Pol.* 48, 1109–1112