Biocybernetics and Biomedical Engineering
2010, Volume 30, Number 3, pp. 35–47

Proteins and Peptides Identification from MS/MS Data in Proteomics

LECH RACZYŃSKI*, TYMON RUBEL, KRZYSZTOF ZAREMBA

Institute of Radioelectronics, Warsaw University of Technology, Warsaw, Poland

Protein identification in biological samples is the most important task in proteomics. In the past decade, mass spectrometry (MS) became the method of choice for the identification of proteins. The purpose of this paper is to give an overview of MS-based protein identification methods, discuss their advantages and limitations and to highlight some recent advancements in this field.

K e y w o r d s: proteomics, tandem mass spectrometry MS/MS, database protein identification, de novo sequencing

1. Introduction

Proteomics is the study of the proteome. The term proteome defines the entire protein complement of the genome. Thus, the objective of proteomics is large-scale analysis of protein function, structure, post-translational modifications, cellular localization and protein-protein interactions. Proteomics is closely related to genomics, but is more complex because, while the genome is constant, the proteome composition varies depending on tissue type, life cycle stage and environmental conditions. As proteins play crucial roles in virtually all biological processes, their dynamically changing expression can be treated as a sensitive indicator of the organism's state. This is why clinical research also may benefit from proteomics by both, the identification of new drug targets and the development of new diagnostic markers. Although many different analytical techniques are used in protein analyses, including one- and two-dimensional gel electrophoresis [1], liquid chromatography [2] and X-ray crystallography [3], current high-throughput proteomic studies mostly rely on mass spectrometry [4], especially in applications related to protein identification. In this

* Correspondence to: Lech Raczyński, Institute of Radioelectronics, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland, e-mail: lraczyn@ire.pw.edu.pl
Received 4 September 2009; accepted 26 March 2010

review, we describe the principles of different strategies for the MS-based protein identification and discuss their strengths and limitations.

2. Peptides and Proteins

Peptides and proteins are biomolecules made of amino acids. All of the 20 naturally occurring amino acids contain a central carbon atom to which a hydrogen atom, the amino (NH_2) and carboxyl (COOH) groups and an amino acid-specific side chain are attached. Peptide is a molecule composed of at least two amino acids held together by the peptide bond between the amino and carboxylic acid groups. The joined amino acids in the peptide are called residues. The first residue in the chain is called N-terminus and the last C-terminus. Although there is no rigid rule, a chain length longer than 30 to 40 amino acids is called a protein. The primary structure (i.e. the amino acid sequence) is a fundamental property of a protein as it determines all other biochemical and biophysical properties. Knowledge of the primary structure is also crucial for protein identification.

3. Mass Spectrometry

Mass spectrometry is an analytical technique based on measurement of mass to charge ratio (*m/z*) of ions. The mass spectrometer consists of three basic elements: an ion source which converts neutral sample molecules into gas phase ions, a mass analyser which separates the ions depending on their *m/z* value and a detector that registers the number of ions of each species. The analysis is carried out under high vacuum, and the whole process is controlled by a computer system which is also responsible for data storage, processing and visualization. The data output is in the form of a mass spectrum, usually is presented as a plot of *m/z* values of the detected ions versus their abundances (Fig. 1). The peaks in the spectrum can represent either intact molecular ions or their smaller fragments.

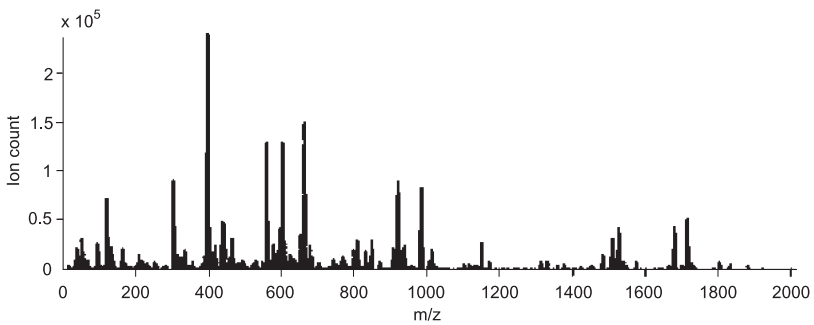

Fig. 1. An example mass spectrum of a complex biological sample. The data was collected on an high-resolution FT-ICR mass spectrometer

Application of mass spectrometry to the analysis of proteins became possible after the development of 'soft ionization' methods, particularly ElectroSpray Ionization (ESI) [5] and Matrix-Assisted Laser Desorption/Ionization (MALDI) [6]. In ESI the solvent containing the analyte is dispersed into a fine aerosol in presence of electric field, while in MALDI laser pulses are used to sublimate and ionize the sample molecules out of a dry, crystalline matrix. Both the techniques allow to transfer large and thermally labile biomolecules into gas phase without leading to their fragmentation. If the analysed molecules have functional groups that readily accept protons, as in proteins and peptides, then positive ionisation is used. The *m/z* value of a protonated molecular ion is calculated according to the formula (1):

$$m/z = \frac{M + n \cdot H}{n}, \tag{1}$$

where M is the molecular mass of the molecule, H is the mass of proton and n is the number of attached protons. Multiply charged ions are usually observed for large molecules when ESI is used, whereas MALDI results predominantly in generation of the singly charged ions.

The performance of a mass spectrometer depends mainly on the capabilities of the used mass analyser. Recent improvements in analysers parameters, such as sensitivity, mass measurement accuracy and resolving power were essential for wide application of the mass spectrometry in the field of proteomics. The four basic types of analysers used in proteomics research are: quadrupole [7], ion trap [8], time-of-flight (TOF) [9] and orbitrap [10]. Less common is the usage of the most powerful, but expensive and technically demanding Fourier Transform Ion Cyclotron Resonance (FT-ICR) [11] analyser.

Determination of peptide sequences requires two steps of mass spectrometry, separated spatially or in time. In the first step, one ion species (the parent or precursor ion) is selected and then is fragmented inside the spectrometer, usually by collision with an inert gas. Next, the resulting fragment ions are registered in the second step, producing a MS/MS spectrum. Instruments allowing such analysis are called tandem mass spectrometers. If the MS stages are separated in space, then the spectrometer must be equipped with at least two analysers. This the case of triple-quadrupole, quadrupole-TOF (Q-TOF) and TOF-TOF instruments. Trapping instruments are capable of performing time-separated tandem spectrometry in a single analyser.

Recently, mass spectrometers are often directly coupled with a High Performance Liquid Chromatography (HPLC) system. In such case, the LC-MS run consists of the repeated MS measurements of the subsequent fractions of the sample eluting from the chromatographic column. Chromatographic separation of the sample compounds prior to the MS analysis results in increased sensitivity and allows to identify analytes which might not be distinguished because of close mass to charge ratio.

4. Protein Identification by Mass Spectrometry

In general, unambiguous identification of intact proteins by mass measurement alone is a difficult task. Therefore, most proteomics experiments are performed according to the 'bottom-up' strategy, which relies on peptide-level information. In this approach proteins are first digested by a specific proteolytic enzyme, and then the peptides are analysed using MS. The most commonly used enzyme is trypsin, which cleaves the sequence on the C-terminal side of arginine and lysine residues, unless the subsequent residue is a proline. The enzyme specificity enables reliable identification based on even a small subset of peptides, covering only partially the full sequence of the protein.

Historically, the first available method of protein identification was Peptide Mass Fingerprint (PMF) [12]. In this technique, proteins are identified by matching the list of measured masses to the list of peptide masses generated by an *in silico* digestion of all entries of a protein sequences database. Unfortunately, this approach is ineffective when applied to protein mixtures, and can be used only in conjunction with prior sample separation by either one- or two-dimensional electrophoresis.

Direct analysis of complex biological samples containing up to thousands of proteins involves the usage of tandem mass spectrometry, in which both intact mass of the peptide and masses of the fragment ions generated inside the spectrometer by the dissociation of the peptide bonds are measured. Tandem mass spectrometry data are usually interpreted in a computer-aided manner, and the currently used algorithms can be divided into three main categories: database searching, *de novo* sequencing and sequence tags. Each of these approaches relies on the knowledge of peptide fragmentation rules, which allows to predict the ions *m/z* values in the MS/MS spectrum of a given amino acid sequence.

4.1. Peptides Fragmentation Rules

A number of peptide fragmentation methods have been developed, including Collision Induced Dissociation (CID) [13], Surface Induced Dissociation (SID) [14], Electron Capture Dissociation (ECD) [15], Electron Transfer Dissociation (ETD)[16] or Infra-Red MultiPhoton Dissociation (IRMPD) [17]. Currently, majority of commercially available tandem spectrometers use low-energy CID, in which the fragmentation is induced by collisions with an inert gas, most commonly argon or helium.

As shown in Fig. 2, in low-energy CID fragmentation mainly occurs along the peptide backbone. There are three different types of bonds that can be dissociated by collisions: the NH-CH, CH-CO, and CO-NH bonds. Each bond breakage produces two ions, and only the charged one is measured by the mass spectrometer. The charge can stay on either of the two sides of the breakage, which means that there are six possible fragment ions for each amino acid residue. The charged fragments containing the N-terminal residue are denoted with letters *a, b, c*, and those containing the

C-terminal residue with letters *x, y, z* [18]. Numerical subscripts of the fragment ions indicate the positions of the amino acid residues at which the bond cleavage have occurred.

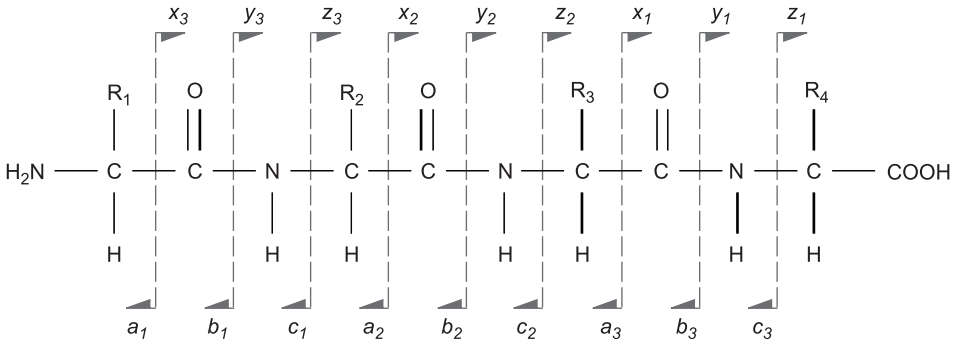

Fig. 2. Schematic representation of the peptide fragmentation process in low-energy CID. R_i denotes the side chain of the ith residue. The arrow directions indicate the charged fragment

Fragment ions generated by dissociation of a single bond give rise to series of peaks in the MS/MS spectrum. Mass differences between adjacent peaks of the series correspond to amino acid residues masses, thus are sequence-specific (Fig. 3). Of the six types of ions, *b* and *y* are formed more frequently, especially for tryptic peptides. Often, *a* ions are also formed along with *b* ions (by the loss of CO from *b*). Additionally, sequence-specific peaks caused by neutral losses from the primary six types of ions

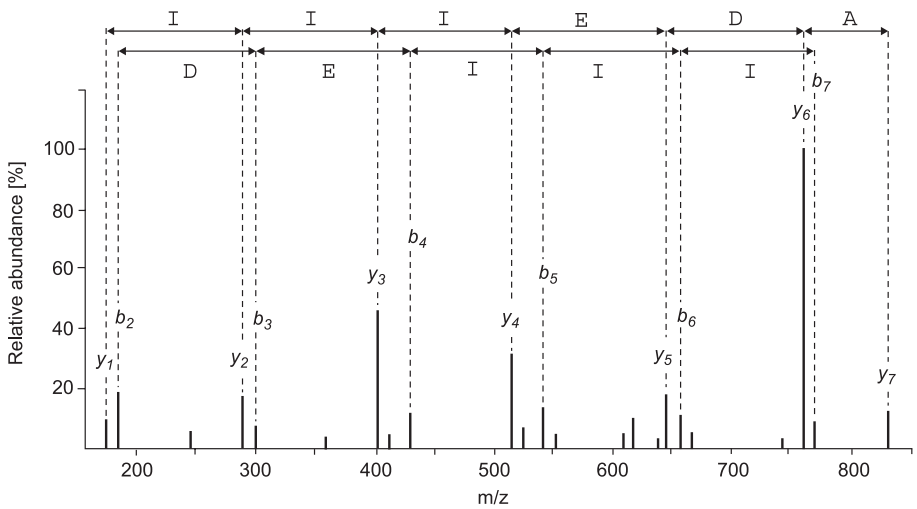

Fig. 3. An example MS/MS spectrum for the peptide LADEIIIR. Sequence identifications based on y and b ion series are equivalent. The C-terminal residue (R) is given by the position of the y_1 ion. The N-terminal residue (L) can by identified on the base of the mass difference between the precursor ion and the peak y_7

can be observed in the spectrum. The most common losses are of ammonia and water from the *b* and *y* ions. In general, ions formed by simultaneous cleavage of two bounds (internal ions) do not carry sequence-specific information. A special case are immonium ions, which are internal fragments with just a single side chain, formed by a combination of *a* type and *y* type cleavage. Peaks of such ions are not sequence-specific, but they can be used to confirm of the existence of residues.

4.2. Database Search Algorithms

Database searching is currently the most widely used method for peptide and protein identification. In this approach, all protein sequences in a database are first *in silico* digested into proteolytic peptides. Next, theoretical MS/MS spectra are computed for peptides with masses falling into a preselected tolerance window around the measured mass of the precursor ion. The experimental spectrum is then compared to the theoretical spectra of database candidate peptides, and the best matching one is selected. The process is repeated for each registered MS/MS spectrum, yielding sequences of the peptides and the corresponding proteins present in the mixture.

A number of database search algorithms have been developed: SEQUEST [19], MASCOT [20], SCOPE [21], PEPHMM [22], OLAV [23], PROBID [24], OMSSA [25]. Different software tools use different criteria to determine similarity between the experimental and the theoretical spectrum. Here two of them will be shortly presented: SEQUEST – the earliest used software for protein identification, and MASCOT – one of the most widely used tool employing the database searching approach.

SEQUEST uses a two-step scoring scheme. The preliminary score restricts the number of analyzed sequences to 500. This score sums the peak intensities of fragment ions matching the predicted sequence ions. It also takes into account continuity of the ion series and the presence of immonium ions. The second score is cross-correlation of the theoretical and the experimental spectrum. The correlation function is calculated according to the formula (2):

$$f(\tau) = \sum_i x_i \cdot y_{i+\tau}, \tag{2}$$

where τ is a displacement value between the two signals. If two signals are the same, the correlation function should maximize at $\tau = 0$, where there is no offset between the signals. The final score attributed to each candidate peptide sequence is value of the function at $\tau = 0$ minus mean of the cross-correlation function over the range $-75 < \tau < 75$. The first ranked spectrum is then assumed to represent the correct sequence. Additionally, the normalized difference between the best score and each of the others, ΔC_n, is produced. This value is important in distinguishing the best sequence from the others. If a match is reasonably unique, this value should be large.

MASCOT considers matches between the peptide from the database and the ions from the MS/MS spectrum as random events. For each peptide, probability that its matches occur randomly is calculated using an empirically generated distribution of the fragment ion probabilities in the database. The probability of the match should be the smallest for the correct amino acid sequence, because most of peaks in the theoretical spectrum will be present in the MS/MS spectrum. Relying on this probability, a s*core* for every peptide is calculated according to the formula (3):

$$Score = -10 \cdot \log_{10}(P) \tag{3}$$

where P is the probability that the observed match is a random event. Because the protein sequences in the database are not random, the best s*core* does not necessarily lead to a correct match. MASCOT offers two statistical thresholds that discriminate between correct and incorrect peptide identifications. First of them, the Mascot Identity Threshold (MIT), is calculated according to the formula (4):

$$MIT = -10 \cdot \log_{10}\left(20 \cdot \frac{\alpha}{N}\right), \tag{4}$$

where α is probability of a random peptide match and N is number of the peptide candidates. As the quality of the measured MS/MS spectrum is not always ideal (more in chapter 4.5), it may not be possible to reach the identity threshold score, even if the best match in the database is a clear outlier from the distribution of random s*cores*. To assist the identification of such outliers Mascot Homology Threshold (MHT) is also reported. This threshold is an empirical measure of whether the match is an outlier from the distribution of *scores* of the candidate peptides. Unfortunately, an exact definition of the MHT was not published by the software authors.

4.3. *De novo* Peptide Sequencing Algorithms

Database algorithms are useful only for identification of the peptides present in a protein database. There are situations when an adequate database is not available or the full sequence of the analyzed protein is unknown so far. In that case, the only way to recognize a peptide is the *de novo* sequencing. It must be stressed, that the *de novo* sequencing is a definitely more difficult task than database searching, as the algorithm tries to predict the amino acid sequence directly from the MS/MS spectrum. This means, that in the *de novo* sequencing all possible sequences may be generated regardless of their biological relevance. To avoid false results, criteria to determine the likelihood function between the candidate peptide and the spectrum peaks must be very restrictive. Thus, the *de novo* sequencing results can be used to validate the database search results. Significant similarity between the result of that two methods could be taken as evidence that the sequence from the database is correct.

As in the case of database methods, there are many *de novo* algorithms: PEAKS [26], PEPNOVO [27], LUTEFISK [28], SHERENGA [29], PROBSEQ [30], NOVO-HMM [31]. Here two of them will be shortly described: SHERENGA – the earliest used *de novo* software, and PROBSEQ – one of the most efficient tools.

SHERENGA, like most *de novo* algorithms, employs the spectrum graph approach for generating of the sequence candidates. This transformation makes it easy to explain the relation between adjacent peaks in MS/MS spectrum. Following the fragmentation rules, if one fragment ion has one more amino acid than another, the *m/z* difference between the two corresponding peaks of the spectrum will be equal to the mass of the amino acid divided by the charge state. The peaks in the spectrum serve as vertices in the spectrum graph, while the edges of the graph correspond to linking vertices differing by the mass of the amino acid. The *de novo* peptide sequencing problem is defined as finding the longest path in the resulting acyclic graph. SHERENGA uses a probabilistic model to evaluate probability that a peptide under consideration may produce an observed spectrum. The program allows the situation in which not all peaks in the spectrum are present. Having the knowledge from the learning dataset, SHERENGA uses 'a premium for explained ions' and 'penalty for unexplained ions' approach.

PROBSEQ has a different approach to obtain the information from the spectrum. In spite of creating the spectrum graph, PROBSEQ creates a probabilistic model based on the Bayesian theory according to the formula (5):

$$\Pr(Seq / Spec) = \frac{\Pr(Seq) \cdot \Pr(Spec / Seq)}{\Pr(Spec)}. \tag{5}$$

The best match is given by the sequence (*Seq*) that maximizes probability on the left-hand side of equation (5). The Pr(*Seq*) is the prior probability of observing a sequence (*Seq*), based on the natural abundances of its constituent amino acids and the preference for C-terminal residues. The Pr(*Spec/Seq*) is probability that the sequence under consideration (*Sec*) could give rise to the observed spectrum (*Spec*). The Pr(*Spec*) is just a normalization variable. In the first step, PROBSEQ generates a population of peptides that are consistent with the intact mass of an unknown peptide. The implementation used in the software does not perform an exhaustive search of an entire space, but simulates it by sampling of possible peptide sequences through a terminated Markov Chain Monte Carlo algorithm. Initially, a trial set of solutions is constructed by sampling from a prior distribution. For those candidate sequences the likelihood calculations are proceeded. Basing on the results, in the next step, the list of peptides is modified by sequence transitions such as: reversal, rotation, permutation or replacement of a contiguous subsequence with randomly chosen end-points. The result of *de novo* exploration is a number of candidate peptide sequences that may account for the fragmentation spectrum and precursor mass.

4.4. Sequence Tags Algorithms

Sequence tagging combines two of the previously described approaches. The first step is the *de novo* sequencing investigation which provides a partial sequence, and then the database searching is done in order to find the full sequence. An example of a sequence tag is [343.12]AGPVIKED[195.66], where the numerical values in the brackets represent masses of amino acid subsequences remaining unidentified because of the lack of information in the MS/MS spectrum. The above tag example is typical, as the fragmentation occurs more often towards the middle of a peptide rather than at its ends. The sequence tagging programs consolidate advantages of both previous methods. The results are validated by the database knowledge of peptides sequences, and on the other hand, generated peptides are selected with extra MS/MS spectrum information. Currently, there are several well known sequence tagging software tools: MS-SHOTGUN [32], FASTS [33], OPENSEA [34], GUTENTAG [35], SPIDER [36] and DeNovoID [37].

4.5. Sources of Failure to Identify Correct Peptide Sequences

There are several reasons why the peptide identification tools fails to assign correct peptide sequences to experimental MS/MS spectra. Limited sensitivity, mass measurement accuracy and resolving power of the used instrument in conjunction which chemical and electronic noise may lead to low quality and incomplete spectra, which cannot be correctly interpreted. Errors in low-level spectra processing, resulting in incorrectly determined charge states or monoisotopic masses of the precursor ions are also common causes of incorrect identifications. Another problem is the fact that some MS/MS spectra originate from non-peptide sample impurities or are result of the simultaneous fragmentation of the different peptide ions having similar *m/z* values.

It can be presumed that improved preprocessing methods [38, 39] and algorithms allowing to remove poor quality spectra from the dataset prior the submission to the peptide identification tool [40, 41] should considerably lower the false identification rates. A more fundamental limitation of the identification process performance emerge from deficiencies in the commonly used scoring schemes and the theoretical fragmentation models. For most peptides, cleavages at the amide bonds are predominant, producing a series of a *b* and *y* ions. However, there are peptides, where the enhanced cleavage occurs at just one or two amino acid residues. Most of the used database search algorithms use simple fragmentation models, based on the assumption that peptide bonds dissociate in an uniform manner, hence, peptides with unusual fragmentation patterns are not sequenced well. To better accommodate these peptides more complex fragmentation models, such as the presented below ‘mobile proton’ model, must be incorporated into the identification tools.

4.6. 'Mobile Proton' Model

The investigation of fragmentation process, performed at variety of conditions, allowed to formulate the 'mobile proton' model [42, 43]. According to the model assumption, the fragmentation of peptides requires the involvement of a proton at the cleavage site i.e. the cleavages are 'charge-directed'. Protonation at the amide bonds (Fig. 4) initiates 'charge-directed' cleavages of the backbone and leads to the *b* and *y* ions generation. Second possibility is that an amino acid side-chain tightly binds a proton (Fig. 4), and the additional energy will be required to move that proton from the basic side-chain to the peptide backbone to induce dissociation. This is the situation when the basic amino acid (arginine, lysine, histidine) is present in a peptide sequence. Location of proton/protons in the peptide depends on the amino acid composition versus the number of protons. If a basic amino acid occurs in peptide sequence, proton will be sequestered by its amine group, so called 'non mobile proton'. In turn, proton bound to N-terminal of peptide will need much lower energy to migrate and induce a dissociation, so called 'mobile proton'. Under high energy dissociation conditions, there are plenty of peaks in spectrum, and fragmentation pattern becomes unpredictable. That way, in typical experiments, low energy collision is used to produce fragment at amide bond generally.

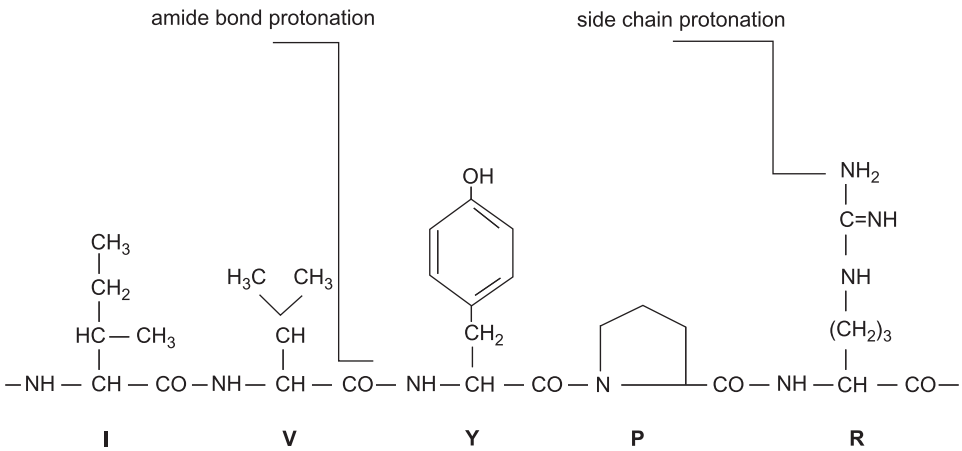

Fig. 4. Protonation sites of a short peptide

The 'Mobile proton' model was used by several groups to create a theoretical spectrum. Shutz et al. [44] derive a new classification scheme, the 'relative proton mobility' scale (RPM), which expands the current 'mobile proton' model. Peptides are classified into three groups. If the number of charges is less than or equal to the number of arginine residues in the peptide sequence, it is classified as 'non-mobile',

'mobile' if the number of charges is greater than the total number of basic residues, otherwise, 'partially mobile'. They built a simple, linear model for predicting ion intensities using the RPM scale. The quality of the predictions varies depending on the peptide charge state giving better results for single-charged peptides than for double-charged peptides. Quality of the prediction varies also according to the mobility state with best results for 'mobile' peptides and the worst for 'non-mobile' peptides.

Zhang [45] followed a different approach and predicted product ion spectra, based on the 'mobile proton' model of peptide fragmentation, using a kinetic model. The kinetic model describes the reaction between molecules involving many competing pathways, such as the case of peptide fragmentation. The model includes most fragmentation pathways described in the literature. The mathematical model is able to predict peptide CID spectra, achieving reasonable accuracy for the fragment ion intensities for both, singly and doubly charged peptide parent ions.

5. Conclusions

Proteomics is a domain of science which is developing very quickly. Its development depends on the possibilities offered by the computer techniques of data analysis and data mining. More effective and powerful peptide identification algorithms are still created. However, the peptides sequencing still remains a difficult problem and subject, which we are unable to solve efficiently. The significant achievement of the last few years has been the model of a fragmentation of a peptide - the 'mobile proton' model. The study of that problem is in progress. Positive results which confirm proposition applied in the model, motivate to put more efforts and work into this subject.

References

1. O'Farrell P. H.: High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 1975, 250, 4007–4021.
2. Witkiewicz Z.: Introduction to Chromatography (in Polish). WNT, Warszawa 1995.
3. Luger P.: Modern X-Ray analysis on single crystals (in Polish). PWN, Warszawa 1989.
4. Hoffman E., Charette J., Stroobant V.: Mass spectrometry (in Polish). WNT, Warszawa, 1998.
5. Fenn J. B., Mann M., Meng C. K., Wong S. F. ,Whitehouse C. M.: Electrospray ionization for mass spectrometry of large biomolecules. Science 1989, 246, 64–71.
6. Karas M., Hillenkamp F.: Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. 1988, 60, 2299–2301.
7. Paul W., Steinwedel H.: A new mass spectrometer without magnetic field. Z. Naturforsch. 1953, 8a, 448–450.
8. Paul W., Reinhard P., Zahn O.: The electric mass filter as mass spectrometer and isotope separator. Z. Phys. 1958, 152, 143–182.

9. Stephens W. E.: A pulsed mass spectrometer with time dispersion. Phys. Rev. 1946, 69, 691.
10. Makarov A.: Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal. Chem. 2000, 72, 1156–1162.
11. Sommer H., Thomas H. A., Hipple J. A.: Measurement of e/m by cyclotron resonance. Phys. Rev. 1951, 82, 697–702.
12. Pappin D. J., Hojrup P., Bleasby A. J.: Rapid identification of proteins by peptide-mass fingerprinting. Current Biology 1993, 3, 327–332.
13. Jennings K. R.: Collision-induced decompositions of aromatic molecular ions. Int. J. Mass Spectrom. Ion Phys. 1968, 1, 227–235.
14. Mabud M. A., Dekrey M. J., Cooks R. G.: Surface-induced dissociation of molecularions. Int. J. Mass Spectrom. Ion Proc. 1985, 67, 285–294.
15. Zubarev R. A., Kelleher N. K., McLafferty F. W.: Electron capture dissociation of multiply charged protein cations: a nonergodic process. J. Am. Chem. Soc. 1998, 120, 3265–3266.
16. Syka J. E. P., Coon J. J., Schroeder M. J., Shabanowitz J., Hunt D. F.: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. USA 2004, 101, 9528–9533.
17. Little D. P., Spier J. P., Senko M. W., O'Conner P. B., McLafferty F. W.: Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. Anal. Chem. 1994, 66, 2809–2815.
18. Roepstorff P., Fohlman J.: Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed. Mass Spectrom. 1984, 11, 601.
19. Eng J. K. et al.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 1994, 5, 976–989.
20. Perkins D. N. et al.: Probability-based protein identification by searching sequence database using mass spectrometry data. Electrophoresis 1999, 20, 3551–3567.
21. Bafna V., Edwards N.: SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 2001, 17, S13–S21.
22. Wan Y. et al.: PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. Recomb. 2005, 342–356.
23. Colinge J. et al.: OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003, 3, 1454–1463.
24. Zhang N. et al.: ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2002, 2, 1406–1412.
25. Geer L. Y. et al.: Open mass spectrometry search algorithm. J. Proteome Res. 2004, 3, 958–964.
26. Ma B. et al.: PEAKS: powerful software for peptide de novo sequencing by MS/MS. Rapid Commun. Mass Spectrom. 2003, 17, 2337–2342.
27. Frank A., Pevzner, P.: Pepnovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem. 2005, 77, 964–973.
28. Taylor J. A., Johnson R.S.: Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal. Chem. 2001, 73, 2594–2604.
29. Dancik V. et al.: De novo peptide sequencing via tandem mass-spectrometry. J. Comput. Biol. 1999, 6, 327–342.
30. Skilling J. et al.: ProbSeq - a fragmentation model for interpretation of electrospray tandem mass spectrometry data. Comp. Func. Genom. 2004, 5, 61–68.
31. Fischer B. et al.: NovoHMM: a hidden Markov model for de novo peptide sequencing. Anal. Chem. 2005, 77, 7265–7273.
32. Huang L. et al.: Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches. J. Biol. Chem. 2001, 276, 28327–28339.
33. Mackey A. J. et al.: Getting more for less: algorithms for rapid protein identification with multiple short peptide sequences. Mol. Cell. Proteomics 2002, 1, 139–147.

34. Searle B.C. et al.: High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal. Chem. 2004, 76, 2220–2230.
35. Tabb D. L., Saraf A. and Yates J. R.: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem. 2003, 75, 6415–6421.
36. Han Y. et al.: SPIDER: software for protein identification from sequence tags with de novo sequencing error. J. Bioinform. Comput. Biol. 2005, 3, 697–716.
37. Halligan B. D. et al.: 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. 2005, 33, 376–381.
38. Mujezinovic N. et al.: Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. Proteomics 2006, 6, 5117–5131.
39. Gentzel M. et al.: Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics 2003, 3, 1597–1610.
40. Moore R. E., Young M. K. and Lee T. D.: Method for screening peptide fragment ion mass spectra prior to database searching. J. Am. Soc. Mass Spectrom. 2000, 11, 422–426.
41. Bern M. et al.: Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 2004, 20, i49–i54.
42. Wysocki V., Tsaprailis G., Smith L., Breci L.: Mobile and localized protons: a framework for understanding peptide dissociation. J. Mass Spectrom. 2000, 35, 1399–1406.
43. Gu C., Somogyi A., Wysocki V., Medzihradszky K.: Fragmentation of protonated oligopeptides XLDVLQ (X=L, H, K or R) by surface induced dissociation: additional evidence for the ‘mobile proton’ model. Analytica Chimica Acta 1999, 397, 247–256.
44. Schutz F., Kapp E. A., Simpson R. J., Speed T. P.: Deriving statistical models for predicting peptide tandem MS product ion intensities. Biochemical Society Transactions 2003, Vol. 31, part 6.
45. Zhang Z.: Prediction of Low-Energy Collision Induced Dissociation Spectra of Peptides. Anal. Chem. 2004, 76, 3908–3922.