De Novo Interpretation of Tandem Mass Spectrometry Data

Subject proposed by Behshad Behzadi

`behzadi[at]lix.polytechnique.fr` *difficulté= moyen (**)*

Votre rapport et votre exposé peuvent être en anglais ou en français.

1 Introduction

The amino acids are the structural blocks of proteins and peptides. There are 20 different amino acids. A peptide is a sequence of amino-acids. Peptide Sequencing via tandem mass spectrometry (MS/MS) is a powerful tool in proteomics for identifying the proteins. Two different approaches are used for this aim. The first approach is searching the genome database and to find the best sequence of the database which matches the spectrum. This method however reliable, in most of cases is not able to interpret the spectra. The second method is the *de novo* spectral interpretation which involves to automatically interpret the spectra using the table of amino acids masses.

2 Problem

The peptide sequencing problem is to derive the sequence of the peptides given their MS/MS spectra. During the mass spectrometry process the peptide sequence with some charges is broken in different positions. Depending on the fact if some charges remain on the fragments the mass/charge ratio of the fragment is observed as peak. For an ideal fragmentation process and ideal mass spectrometer the sequence of a peptide could be simply determined by converting the mass differences of the consecutive ions in a spectrum corresponding to amino acids. In practice, the experimentations are far from the ideal case. Thus a de novo algorithm can provide valuable information about the spectrum.

The input of the problem consists of three parts:

- The charge of the peptide (basically the charge is 1, 2 or 3).
- The mass/charge ratio of the peptide.

- The spectrum which is defined by list of ordered pairs of form (x, y). x corresponds to the mass/charge ratio of a fragment (typically a charged suffix or prefix of the peptide or a noise peak) and y is the intensity obtained for this m/z.

The output should be the best prediction of the peptide sequence. If the sequence cannot be identified completely the partial reliable subsequences can be reported. It would be interesting to have a list of candidate peptides with probabilities. For a noise-free complete input, the output should be the complete peptide sequence.

3 Example

Here we present an example of a spectrum which corresponds to a doubly charged peptide. Note that when the peptide is doubly charged the fragments are either doubly charged or singly charged. The input of the problem as stated before is the m/z value 680, the charge 2 and the given spectrum while the output should be the sequence (or some subsequences the sequence) given in the upper part of the figure (`YTGAGMNPARSFA`). Note that the mass of this peptide is approximately 680.0*2=1360.

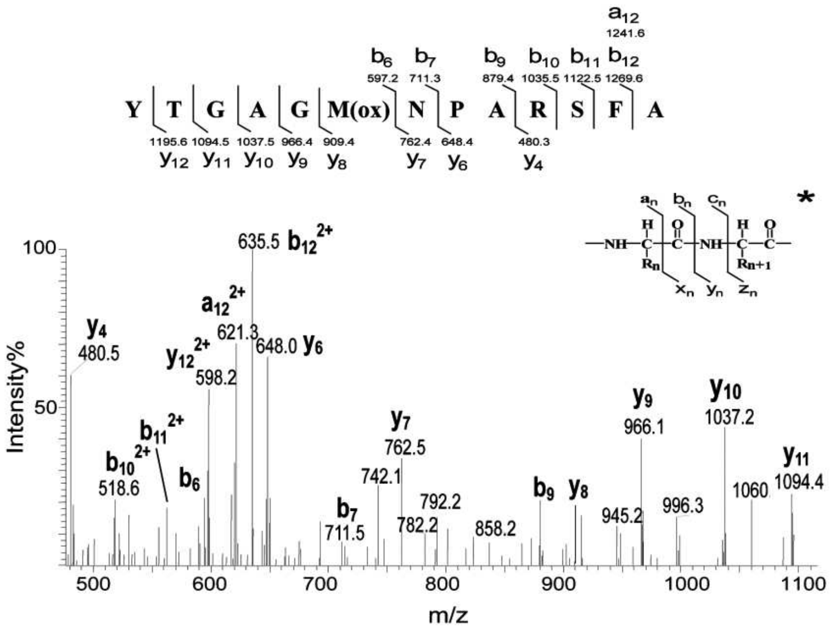

Figure 1: A typical spectrum for a peptide of m/z 680.0

Let us give evidences that how on can construct the sequence `YTGAGMNPARSFA` can be found from the given spectrum. The differences of the x coordinate of the three high consecutive peaks at 879.4, 1035.5 and 1122.5 (156.1 and 87.0) denote approximately the masses of the amino acids R and S respectively (see the url of the masses of amino acids presented in the next section). This shows that the substring RS is probably a substring (a tag) of the sequence. Different tags can be found in the similar way. Note that the tags can have different orientations. One way to find the best peptide sequence is to generate all the possible sequences and to choose the sequence which has the higest score. The score can for example be defined by the number of matched peaks or the length of the matched tags. This way of computation is too expensive in the terms of execution time. Dynamic programming and graph theoretical approaches have been propoed for computing efficiently a reliable sequence.

4 Algorithms and Resources

Different algorithms mainly based on graph theoretical approach and dynamic programming have been proposed. *SHRENGA* [1], *PEAKS* [2] and *LUTEFISK* [3] are some of the known ones. An ideal project would be a project which considers the advantages of the three methods and takes in consideration the intensities of the ions as well.

Examples of input mass spectra with the real peptide sequences and the output of the different algorithms for these spectra can be found at `http://www.csd.uwo.ca/~bma/peaks/`. The url `http://haven.isb-sib.ch/tools/isotopident/htdocs/aa-list.html` contains the table of masses of amino-acids.

Bibliographie

[1] V. Dancik, T. A. Addona, K. R. Clauser, P. A. Vath and J. E. Pevner. *De novo* peptide sequencing via tandem mass spectrometry: a graph theoretical approach. *Journal of computational biology*, 6:327–342, 1999.

[2] B. Ma, K. Zhang, G. Lajoie, A. Doherty-Kirby, C. Liang and M. Li. Peaks: powerful software for peptide de novo sequencing by tandemm mass spectrometry. *Rapid Commun. Mass Spectrom.*, 2002.

[3] J.A. Taylor and R. S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandemm mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11:1067–1075, 1997.