

Improving the *de novo* Sequencing Accuracy by Combining Two Independent Scoring Functions in PEAKS Software

Bin Ma¹; Gilles Lajoie²

¹Department of Computer Science; ²Department of Biochemistry, University of Western Ontario, London, ON, Canada

Overview

By combining the original PEAKS scoring function and a new scoring function, the accuracy of PEAKS' *de novo* sequencing is remarkably improved.

Introduction

De novo sequencing from MS/MS data is a standard method for peptide sequencing that does not require the sequences to be in a database, and therefore is best for novel proteins. *De novo* sequencing is also better at finding PTMs. In addition, when homologues of the novel proteins are in the database, they can be found by sequence homology search after *de novo* sequencing. Even for proteins in a database, if *de novo* sequencing computes the correct sequence without looking at the database, the confidence is much higher than simply finding the sequence from the database.

A *de novo* sequencing program typically requires a scoring function that evaluates the fitness between a peptide sequence and the spectrum. The choice of scoring functions affects the program's accuracy significantly. In this abstract we demonstrate that better accuracy can be achieved by combining two independent scoring functions.

Methods

The original PEAKS scoring function¹ is based on the logarithm intensity of all matched peaks. When a y ion is matched (Fig. 1), the score is computed by the following formula, where $f()$ is an empirical function that is in favour of the coexistence of y ion and its neutral loss ions.

$$\log(\text{intensity}) \times e^{-(\text{error}/\text{tolerance})^2} \times f(y-17, y-18, x)$$

The scores of other ion types are also computed similarly. And finally, the total score of all ion types is used to select the best *de novo* sequence.

We recently found that the *rank* of a peak (the number of peaks that are equal to or higher than the peak of interest) is as important as the peak's intensity. For example, as shown in Fig. 2, the y7 ion is very low-intensity but it is still one of the top-50 peaks in the spectrum. If only the intensity is considered, the y7 peak may be determined as a noise peak, but the rank can provide additional proof for this peak to be a real signal.

Using many MS/MS spectra with known peptide sequences, we counted the frequency that each ion type is assigned by a certain rank value, and recorded all the frequencies in a table. When *de novo* sequencing is performed, if the k -th highest peak is matched by an ion, by looking up the table we can find out how often this event occurs, and assign the score according to the frequency. The peptide score is defined to be the sum of the rank scores of all matched peaks. Finally, the new peptide score is added up with the original PEAKS score.

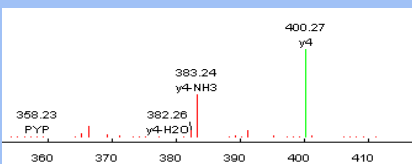


Fig 1. The presence of related ions is recognized in the interpretation

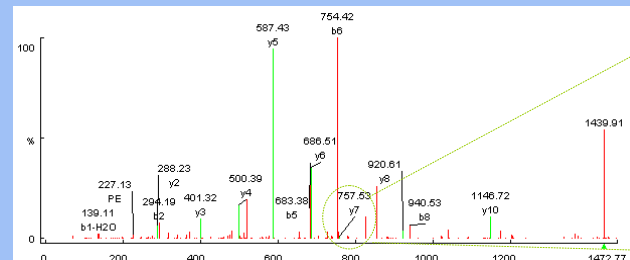
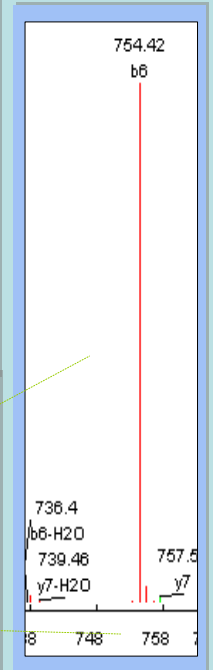
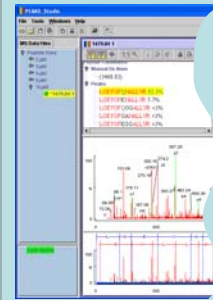


Fig 2. y7 ion is low intensity; but still top-50 in the spectrum.



Experimental Results

The version 2.4 and later versions of PEAKS *de novo* sequencing software combined the original PEAKS score and the new rank score. In contrast, PEAKS 2.3 used only the original score. The accuracies of the two versions were compared by using 61 Q-ToF spectra from two standard proteins BSA_BOVIN and ADH_YEAST.

The 61 spectra were measured with a Q-TOF GLOBAL. A low filter (i.e. 10 cts/sec above background for the precursor ions) was used in the data collection and therefore the quality of several spectra is very low. We only kept the 61 spectra that have at least three strong y-ion matches with some peptides of the two proteins. We noted that three strong y-ion matches are not sufficient for *de novo* sequencing. That is, this data set is a difficult one.

Then PEAKS 2.3, PEAKS 2.4, and another commercially available software tool were employed to compute the sequences *de novo*. The default Q-ToF parameters were selected and PTM was turned off in the software. Three criteria were considered to evaluate the accuracy: (1) correct amino acids (2) completely correct sequences, (3) partially correct sequences with five or more contiguous correct amino acids. The results were shown in table 1.

	Typical other software	PEAKS 2.3	PEAKS 2.4
Correct sequence	7	13	23
Correct tags of length >= 5	24	38	50
Correct AA	233	457	559

Table 1. Comparison on 61 sequences (764 AA in total).

Reference:

1. B. Ma, K. Zhang, A. Doherty-Kirby, C. Hendrie, C. Liang, M. Li and G. Lajoie, *Rapid Communications in Mass Spectrometry* 17(20): 2337-2342. 2003.