Improved positional confidence score in MS/MS peptide *de novo* sequencing

Bin Ma[1]; Gilles Lajoie[2]

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

Overview

A new "positional confidence score" is developed to indicate which parts of the *de novo* sequencing results are correct.

Introduction

De novo sequencing from MS/MS data is used widely for peptide and protein identification. However, due to the imperfections of the data and/or software, the results are not always reliable. Very often, only partially correct sequences can be obtained by *de novo* sequencing. If the correct portions of the sequences are known, they can be used as sequence tags to identify the proteins through a homology search. It is therefore very useful for *de novo* sequencing software to give a positional confidence for each individual amino acid in the peptide it computes from the MS/MS data. We describe here a new method to perform this task.

Method

For each spectrum, the software PEAKS (Ma *et. al*.2003, Rapid Comm. Mass Spectrom. 17(20): 2337-42) is used to generate a complete peptide sequence ***P*** and thousands of additional top-scoring sequences in approximately one second. The additional sequences are sorted in the order of their scores. The score of the *de novo* sequence ***P*** should be no less than the scores of the additional sequences, as illustrated in Figure 1.

In order to estimate the confidence of a given amino acid in ***P***, such as the red letter V in Figure 1, the software finds the first additional sequence that does not agree with ***P*** on the given amino acid. (In Figure 1, this is the sequence with score 30.) If the score of this sequence is significantly lower than the score of ***P***, it means the given amino acid is important for high scoring. Consequently, the given amino acid is likely to be correct. Thus, this "score gap" is used here to assign a confidence to the amino acids.

Results

Three MS/MS datasets were used for testing purposes. The first dataset consists of 144 MS/MS spectra recorded on an LCQ ion trap for tryptic digest of a mixture of several proteins. The second dataset consists of 61 MS/MS spectra measured with a Q-TOF using two proteins, BSA and ADH. The third dataset consists of 22 MS/MS spectra obtained on an Orbitrap using Bovin Lactoglobulin Beta. For the purpose of the testing the *de novo* sequencing, output results must contain errors. Therefore, all the three datasets used here include some spectra with quality lower than typically needed for *de novo* sequencing.

For each amino acid in a *de novo* sequence, we calculate the "score gap" as the confidence score. For each confidence level, the number of correct and incorrect amino acids are counted and plotted in Figure 2 using different colors. As shown in the figure, the confidence score developed here can clearly distinguish the correct amino acids from the incorrect ones. In addition, the figure indicates that the score works better with better data (instrument type).

Discussion and Availability

The "score gap" demonstrates great ability to distinguish the correct and incorrect amino acids in *de novo* sequencing result. Although not given in full details here, with rather simple statistics, the score gap can be converted to a positional confidence score (correctness probability) that ranges between 0 and 1. The PEAKS software already has some built-in positional confidence score. However, future versions of PEAKS will adopt the positional confidence score proposed here.

	sequence	seq. score in PEAKS	score gap
de novo seq.	YV**V**GTHR	47	
additional sequences	YVVGHEK	46	1
	YV**E**GAHR	30	17
	YVGHTR	29	18
	YVEGHVK	29	18
	YVVGEHK	28	19
	YVEGVHK	27	20
	.....	.....	.....

Figure 1. An example of score gap calculation.

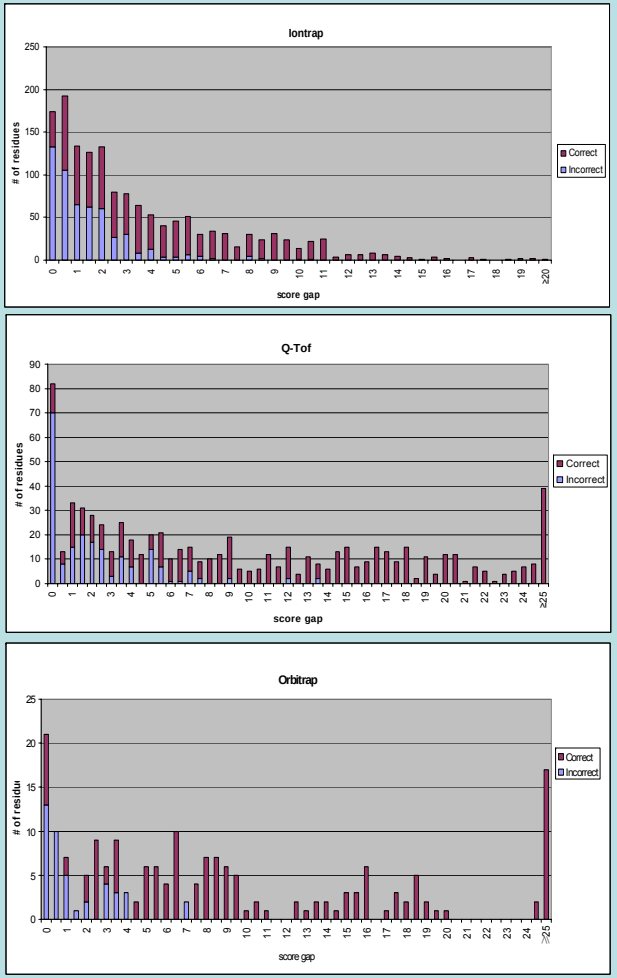

Figure 2. The number of correct and incorrect amino acids at each confidence score value.