No News is Good News: *de novo* Determination of Amino Acids when Peaks are Missing

Lin He and Bin Ma,
David R. Cheriton School of Computer Science, University of Waterloo

Bioinformatics Research Group @ UWaterloo

Overview

Purpose: To improve the accuracy of *de novo* sequencing through the determination of amino acids when peaks are missing.

Methods: The probability that the fragment ions between a pair of amino acids are missing is learned from the NIST database. This probability is used to determine the local sequence when peaks are missing.

Results: Increased the prediction correctness of *de novo* sequencing.

Introduction

Proteomics analysis frequently requires the use of *de novo* sequencing to determine novel peptide sequences. *De novo* sequencing typically requires higher MS/MS data quality than the database search approach. This is because when some fragment ion peaks are missing, a portion of the peptide sequence may become ambiguous, and a *de novo* sequencing software will have to "guess" for that portion of the sequence. This becomes the main reason for the errors made by *de novo* sequencing software. We present a method to significantly increase the success rate of the *de novo* sequencing when some fragment ion peaks are missing.

Methods

The main idea is that certain local sequences tend to lose peaks more often than others. This different tendency significantly helps to derive which sequence is more likely to be correct when peaks are missing.

For every amino acid pair X_1 and X_2, the NIST peptide MS/MS database is used to learn the ion-missing probability between X_1 and X_2 in tryptic peptides. When there are ambiguities for determining the amino acid sequence for a gap in the ion ladders, we use the above probabilities and Bayesian rules to calculate the likelihood of each candidate witch is pre-calculated in a mass table, and output the most likely candidate.

For all the testing peptides, we focus on a type of specific regions where peaks are missing from ion ladders. More specifically, each examined region consists of a 4-mer (four consecutive amino acids) $X_1X_2X_3X_4$, such that both the *b* and *y* ions between X_2 and X_3 are missing, but there are peaks between X_1 and X_2, and between X_3 and X_4, respectively. Thus, a *de novo* sequencing software package would have to guess the two amino acids X_2 and X_3 from a list of candidates with mass equal to the total mass of X_2 and X_3. Suppose there are *k* candidates, then the probability that a random guess is correct should be 1/*k*. For our method, we calculate the likelihood of each candidate and output the most likely one. If there are multiple candidates with the identical likelihood value, then we randomly select one of them and output.

Figure 1 shows an example of such a specific region. During the *de novo* sequencing, D and L can be determined by the first and the last gap appearing in the spectrum, however, the gap in the middle, between peaks at 866.2 and 1070.3 (the denoted mass is 204.1 Da), is not a mass value of any amino acid so that the exact amino acids can not be determined only according the current spectrum. Checking in the pre-calculated mass table using this 204.1 Da, four amino acid pairs, CT, TC, GF and FG, can be chosen as the candidates to determine the local sequence of this spectrum.

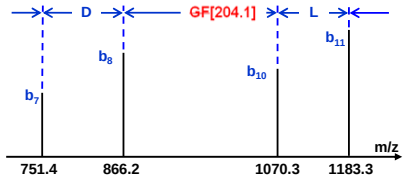

Figure 1 An example of ambiguity during the *de novo* sequencing. V can be determined by the gap between b_7 and b_8, as well D can be obtained by the gap between b_{10} and b_{11}.

Results

168,377 peptides and their MS/MS spectra are used in the experiment. These peptide-spectrum pairs are divided into three dataset according the precursor charge from charge 1 to 3. For each dataset, approximately half of the peptide-spectrum pairs are selected as the training data for learning the ion-missing probability between a pair of amino acids, and the other half as the testing data. All the data are obtained from ion trap instruments and CID is used for the peptide fragmentation.

For MS/MS using CID fragmentation method, *b* and *y* ions are the most commonly observed ions. In our experiment, both types of ions were considered singly and together, respectively.

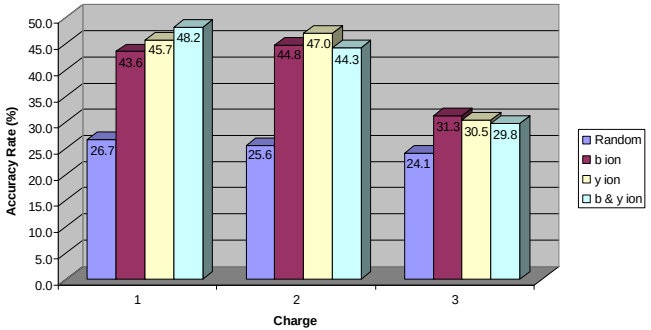

Figure 2 Performance comparison in charge 1, 2 and 3 testing dataset. X axis denotes the charges of three dataset, Y axis denotes the accuracy rate of the prediction.

Figure 2 shows the performance comparison of our method used on three testing dataset which charge 1 to 3. In each testing dataset, four methods were used to predict the amino acids with ambiguities. The first method randomly "guessed" the correct amino acids from the given candidates. The other methods were three instances considering different type of fragment ions. The second and the third method respectively utilized *b* and *y* ions to do the prediction, while the last method considered both *b* and *y* ions. For example, in the charge 2 data, 7,643 4-mers in which the middle peaks missed were used for testing. The random guessing strategy would produce 1,953 (25.5%) correct predictions on the testing cases. By using b-ion alone in our method, the correct prediction was improved to 3,422 (44.8%) cases. Using y-ion alone gives 3,595 (47.0%). The combination of *b*- and *y*-ions together gave 3383 (44.3%). We conjecture that the accuracy can be further improved by combining CID and ETD fragmentations together and the research on this is currently on going.

Conclusion

The experiment results showed that our Bayesian model can significantly increase the accuracy rate of the prediction compared to the commonly used random guessing. This method could help to improve the accuracy of the de novo sequencing when some peaks are missing.