New Method for the Validation of de novo Sequencing Results

Introduction
De novo sequencing from MS/MS data is essential to identify peptides of unknown genomes. Imperfect data quality causes errors in de novo sequencing results. Therefore it is important to have a scoring function that reflects the correctness probability of each de novo sequencing result. Moreover, a de novo sequencing error is usually only partially correct. Hence another score function should produce the correctness of each individual amino acid in a peptide sequence obtained by de novo sequencing.
Since de novo sequencing does not depend on protein databases, the validation and confidence methods developed in the database search approach such as the reverse-database query cannot be applied. Here we present a general validation algorithm which uses any de novo score to calculate the correctness probabilities of each amino acid in the de novo sequencing results. In addition to result validation, these probabilities can also be used in other protein identification software such as Tagger [1].

In this research we demonstrate the method by using PEAKS [3] de novo sequencing score and results. However the method in general enough to be adopted in other de novo sequencing software too.

Methods
In our previous research [2], we proposed the “score gap” idea for predicting the correctness of an amino acid in a de novo sequence, and observed the strong correlation between the score gap and the correctness of the amino acid. Therefore it is important to have a scoring function that re
duces the false detection of a correction amino acid. Hence an ideal score function should predict the correctness of an individual amino acid in a peptide sequence obtained by de novo sequencing.

For a particular amino acid X in peptide P, we can mutate P by changing X together with a few other neighboring amino acids without changing the precursor ion mass. All such mutated peptides are generated and their scores are computed by changing X. The difference is calculated and is called the score gap of X. A table is displayed in Figure 1 to show the score gap caused by mutation.

Distribution of Score Gaps
The assumption is that for a given amino acid X, if X is correct, a random mutation of X will cause a significant drop in the score of the whole peptide i.e., a large score gap. If X is just a random match to the spectrum, then a mutation of X will only cause a minor drop or even raise the score of the whole peptide. Thus score gap can be used to discriminate correct (+) from incorrect (-) amino acids. ROC curves in Fig 2 shows the discriminating capacity of score gaps. Distribution curves for correct (+) and incorrect (-) amino acids are learned from a training set separately. We noticed that different instruments have different distribution curves. Therefore we built a training set for each type of instrument and train distribution curves separately. Two pairs of distribution curves for QTOF and Ion trap data are illustrated in Fig 3a & 3b.

Computation of the Probability Associated with Score Gaps
Once we have the distribution curves and the percentage of correct amino acids p(+) we can compute the probability associated with score gap using the Bayesian formula [4].

\[
p(\text{score gap} | +) = \frac{p(+) \cdot p(\text{score gap} | +)}{p(+)}
\]

In reality, even the same type of instrument will produce data sets of different quality. This usually will cause the distribution curves to shift along with the horizontal axis. To deal with this situation, a standard statistical technique called EM (expectation maximization) [5] is adopted. Before the Bayesian formula is applied, EM will automatically adjust the shape and position of distribution curves according to the current data set.

Experimental Results
Two test datasets, one from an LCQ-IonTrap and the other from Waters QTOF, were used with the PEAKS software for the de novo sequencing. The standard distributions of the score gaps of correct and incorrect amino acids were trained using an unrelaxed set of MS data obtained from the same type of instruments.

The amino acids were binned into groups of approximately 100 according to the assigned p(+) (ScoreGap) probabilities to test the model across a variety of probability cutoffs. In each group, the actual probability was calculated based on the number of correct amino acids relative to the total bin size. The actual probability versus the average calculated p(+) (ScoreGap) probability in each group is shown in Fig 4. In this plot, perfect agreement between the actual probability and the calculated probability would fall on a straight 45-degree line.

Conclusion
A scoring method for measuring the correctness probability of each individual amino acid in de novo sequencing result is given. Experimental results showed excellent agreement between the calculated correctness probability and the real correctness probability. The method we present here can be adapted to different de novo sequencing software with different scoring systems.

Reference
1. Denis Yuan, Bin Ma. Novel scoring function improves homology searches using MS/MS de novo sequencing results (ASMS 2008 poster presentation).
2. Bin Ma, Gilles Lajoie. Improved positional confidence score in MS/MS peptide de novo sequencing (ASMS 2006 poster presentation).

Table 1: Sequence Mutations

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Sequence</th>
<th>Peptide Mass</th>
<th>Score Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation 1</td>
<td>YVVK</td>
<td>812.4181</td>
<td>28</td>
</tr>
<tr>
<td>Mutation 2</td>
<td>YVVK</td>
<td>812.4181</td>
<td>29</td>
</tr>
<tr>
<td>Mutation 3</td>
<td>YVVK</td>
<td>812.4181</td>
<td>19</td>
</tr>
<tr>
<td>Mutation 4</td>
<td>YVVK</td>
<td>812.4181</td>
<td>27</td>
</tr>
<tr>
<td>Mutation 5</td>
<td>YVVK</td>
<td>812.4181</td>
<td>27</td>
</tr>
<tr>
<td>Mutation 6</td>
<td>YVVK</td>
<td>812.4181</td>
<td>27</td>
</tr>
</tbody>
</table>

Figure 1: Score Gap Distribution for LCQ-IonTrap Data

Figure 2: Discriminating Capacity of Score Gaps

Figure 3a: Score Gap Distribution for QTOF Data

Figure 3b: Average Probability Estimation

Figure 4: Accuracy of Probability Estimation