New Method for the Validation of *de novo* Sequencing Results

Lei Xin[1], Gilles Lajoie[2], Bin Ma[2]
[1]Bioinformatics Solutions Inc., Waterloo, ON
[2]University of Western Ontario, London, ON

Bioinformatics Solutions Inc.

Introduction

De novo sequencing from MS/MS data is essential to identify peptides of unknown genomes. Imperfect data quality causes errors in the *de novo* sequencing results. Therefore it is important to have a scoring function that reflects the correctness probability of each sequencing result. Moreover, a *de novo* sequence is usually only partially correct. Hence an ideal score function should predict 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* sequencing scores 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 SPIDER [1].

In this research we demonstrate the method by using PEAKS [3] *de novo* sequencing score and results. However the method is 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 amino acid correctness. In this research we combine the score gap idea with statistical methods to compute a percentage value for the correctness of each amino acid.

Score Gap Computation.

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 calculated using the same score function of the *de novo* algorithm. The highest score is compared with the score of the original peptide P. 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.

Table 1: Sequence Mutations

	Sequences	Mass	PEAKS score	Score gap
De novo sequence	YVVGTHR	812.4293	47	
Mutation 1	YVVGHEK	812.4181	46	1
Mutation 2	YVEGAHR	812.3929	30	17
Mutation 3	YVVGHTR	812.4293	29	18
Mutation 4	YVEGHVK	812.4810	29	18
Mutation 5	YVVGEHK	812.4181	28	19
Mutation 6	YVEGVHK	812.4180	27	20

In order to compute the score gap of amino acid H in the de novo sequence, mutations were created around the amino acid, without changing the peptide mass. The smallest difference between the scores of the mutations and the score of the de novo sequence is used as the score gap of H.

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 may 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 often 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.

Figure 2: Discriminating Capacity of Score Gaps

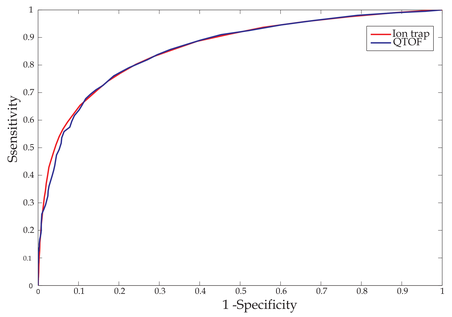


Figure 3a:
Score Gap Distribution for LCQ-ionTrap Data

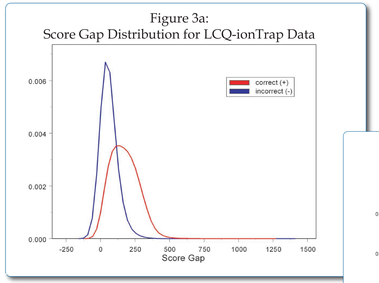
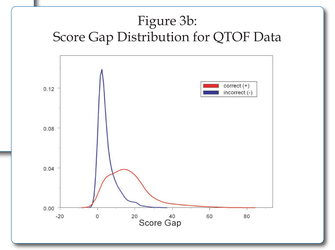


Figure 3b:
Score Gap Distribution for QTOF Data



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]:

$$p(+ \mid ScoreGap) = \frac{p(ScoreGap \mid +)p(+)}{p(ScoreGap \mid +)p(+) + p(ScoreGap \mid -)p(-)}$$

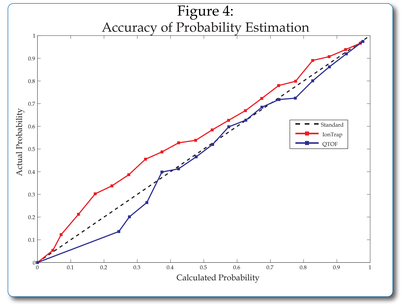
In reality, even the same type of instrument will produce data sets of different quality. This usually will cause the distribution curves to shift along with the horizontal axis. To deal with this situation, a standard statistical procedure called EM (expectation maximization) [5] is adopted. Before 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 unrelated set of MS data obtained from the same types 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.

Figure 4:
Accuracy of Probability Estimation



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 Yuen, 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).
3. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, Gilles Lajoie. PEAKS: Powerful Software for Peptide *De Novo* Sequencing by MS/MS. Rapid Communications in Mass Spectrometry, 17(20):2337-2342. 2003. Early version appeared in 50th ASMS Conference 2002.
4. Devinder S Silva. Data Analysis: A Bayesian Tutorial. Oxford University Press Inc. New York, NY, 1996.
5. Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977