Determining the False Discovery Rate for Peptide Identification without a Decoy Database

Bioinformatics Solutions Inc.

Lei Xin[1], Paul Shan[1], Bin Ma[2]

[1] Bioinformatics Solutions Inc, Waterloo, ON

[2] University of Waterloo, Waterloo, ON

Overview

Purpose: To develop a method to assess the false discovery rate (FDR) of peptide identification without using a decoy database.

Methods: Use the "second best matches" of the spectra on the target database to learn the distribution of the false matches and use that distribution to estimate the FDR of the first matches.

Results: The error of the estimated FDR and the real FDR are usually within +/- 2%.

Introduction

In software peptide identification with MS/MS, being able to estimate the FDR (false discovery rate) of the results is of crucial importance. A popular method today for FDR estimation is to run the search on both the target and decoy databases. This inevitably increases the search time. In addition, it is an elusive problem to generate a decoy database whose distribution is the same as the target database. To solve these problems, we propose to use the "second best matches" of the spectra on the target database to learn the distribution of the false matches and use that distribution to estimate the FDR of the first matches. This method showed excellent performance without any penalty on searching speed.

Methods

First, a more accurate score based on linear discrimination function (LDF) is developed by using four features of the peptide matching: the raw ion matching score, the raw score average for a spectrum, the peptide length and the de novo sequencing score. These four features can effectively discriminate correct matches from false matches.

Secondly, the LDF score distribution of the false matches are learned based on the score of the second best match of each spectrum. This is different from the traditional approach that uses a decoy database. Our method is valid because the second best matches are mostly false discoveries, yet are from the same database. Once we know the LDF score distributions of false matches (-), then the FDR at score threshold is calculated as:

$$FDR = P(-\mid LDF > t) = \frac{P(LDF > t \mid -)P(-)}{P(LDF > t)}$$

Expectation maximization is used to estimate the prior probability:

Figure 1. Raw Score vs. de Novo Score

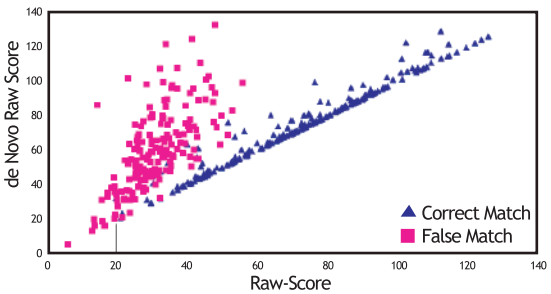


Figure 2. Raw Score vs. Peptide Length

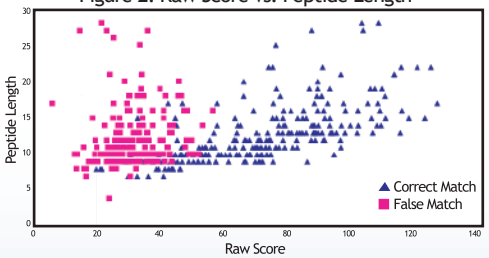
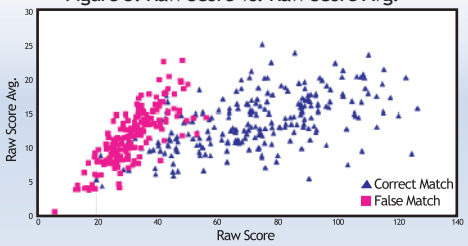


Figure 3. Raw Score vs. Raw Score Avg.



Results

This method was tested by a protein mixture composed of 49 human proteins. The sample was reduced and alkylated by iodoacetamide, then digested by trypsin. Four MS/MS data sets were obtained from four different instruments: LTQ-Orbitrap, Q-TOF, TOF-TOF and Ion-trap. PEAKS® 5.1 was used to identify the peptides from the dataset using NCBI nr protein database.

Figures 1 - 3 show the scatter plots for these four features: the raw ion matching score, the raw score average for a spectrum, the peptide length and the de novo sequencing score.

We compared the FDR calculated using the "second best match" distribution with the real FDR in Figures 4 -7. The real false discovery rate is calculated as the number of false matches above a threshold divided by the total number of peptide matches above that threshold. We can see for most cases the estimated FDR curve matched the real FDR curve pretty well and for TOF-TOF data set they almost overlapped. The error between estimated FDR and real FDR is mostly within +/-5%. When the real FDR is below 5%, the error between the estimated and real FDR is usually within +/-2%.

Figure 4. FDR comparison for Ion-trap data

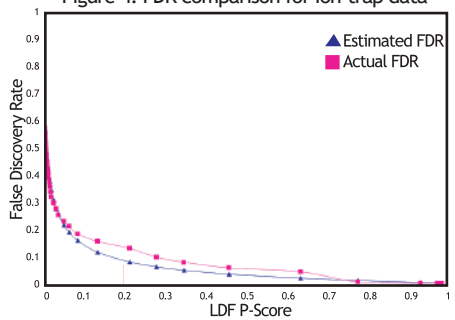


Figure 5. FDR comparison for QTOF data

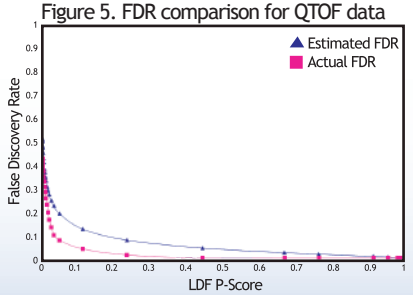


Figure 6. FDR comparison for TOF-TOF data

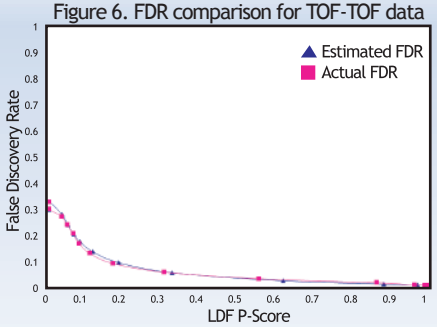
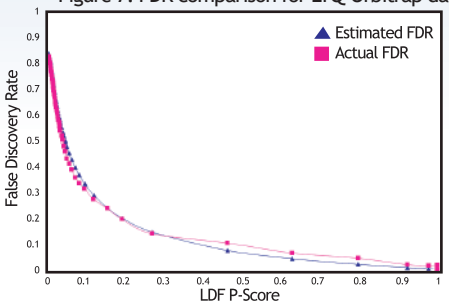


Figure 7. FDR comparison for LTQ-Orbitrap data



Conclusion

This has demonstrated to be a promising method used to estimate the false discovery rate without using a decoy database.

Reference

B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. 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.

G. McLachlan, T. Krishnan. *The EM Algorithm and Extensions*. John Wiley And Sons, New York, 1996.

T. Joachims. *Estimating the Generalization Performance of a SVM Efficiently*. Proceedings of the International Conference on Machine Learning, Morgan Kaufman, 2000.

P. Andrews, D. Arnott, M.A. Gawinowicz. *ABRF-sPRG2006 Study: A Proteomics Standard* (ABRF 2006 poster).