Estimation and Optimization of the Accuracy of Peptide Identifications Obtained by MS/MS Database Searching

Poster Number: MPB 022

Wilfred H. Tang, Sean L. Seymour, Sean P. Keating, Ignat V. Shilov, Alpesh A. Patel, Christie L. Hunter, Daniel A. Schaeffer, Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404, USA

ABSTRACT

We present a methodology for (1) optimizing the discriminating power between correct and incorrect peptide identifications, and (2) estimating the confidence that a peptide identification is correct. The use of this methodology is illustrated by applying it to the Interrogator™ algorithm for database searching. We find that for the Interrogator algorithm, the percent confidence predicted by performing a least-squares linear fit to each MS/MS spectrum's cumulative score distribution provides both good discriminating power and reasonably accurate confidences.

INTRODUCTION

There are now many algorithms available for performing MS/MS database searching. All of these algorithms generally calculate some kind of "score," which measures how closely peptide sequences match MS/MS fragmentation spectra. In a typical peptide identification run, for each experimentally-derived MS/MS spectrum, the algorithm compiles a list of peptide-to-spectrum matching scores. These lists are then presented to the scientist, who must then assess which of the potential peptide identifications on the lists are valid – that is, which peptides are in the actual sample injected into the mass spectrometer. Assessing the validity of potential peptide identifications on the basis of score alone can be difficult and time-consuming. This problem must be addressed before any sort of high-throughput proteomics can be achieved, and several recent publications[1-3] have discussed various aspects of this issue. Here, we present a methodology for (1) optimizing the discriminating power between correct and incorrect peptide identifications, and (2) estimating confidences: probabilities that potential peptide identifications are correct.

MATERIALS AND METHODS

A well-characterized protein mixture was denatured, reduced with dithiothreitol, alkylated with iodoacetic acid, digested with trypsin, separated by reverse-phase liquid chromatography, and injected via electrospray ionization into an API QSTAR® Pulsar LC/MS/MS System. The collected MS/MS spectra were scored by the Interrogator algorithm for database searching. For each MS/MS spectrum, the Interrogator algorithm calculated a score distribution (histogram) – for each peptide in the database which satisfied the error tolerances, a score was calculated by matching a weighted subset of the experimental MS/MS peaks against the theoretical MS/MS peaks expected for that peptide. The resulting score distributions were stored for later analysis and visualization by linear discriminant analysis (LDA)[4], receiver operating characteristic (ROC) plots[5], and least-squares modeling.[6]

Knowing the composition of our experimental sample facilitated annotation of the MS/MS spectra – that is, for each MS/MS spectrum, we determined which potential peptide identification was correct and which was incorrect. (Note: the annotation process was not trivial; for further details, see reference 7.) This annotation was used in data analysis as well as in evaluating the effectiveness of various data analysis methods.

OPTIMIZING DISCRIMINATING POWER

Linear discriminant analysis (LDA) is used to consider the discriminating power of several metrics as well as linear combinations of these metrics. Some metrics considered are:

- High score – for each MS/MS spectrum, the highest score from the score distribution
- Distance-to-pack – for each MS/MS spectrum, the difference between the highest score and the seventh highest score from the score distribution
- Delta mass – the difference between the actual peptide mass and the mass measured by the mass spectrometer

The conclusion obtained by LDA is that distance-to-pack *alone* is the most discriminating metric. The ROC plots below visualize this result. (Note that the ROC plots do not prove the result; it is the LDA which proves the result.)

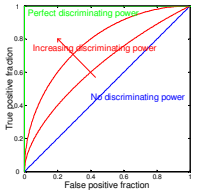

Figure 1. Visualization technique: ROC plots
ROC plots provide a method for visualizing the discriminating power of different metrics.

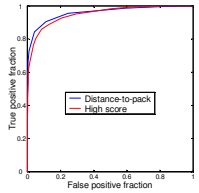

Figure 2. ROC plot comparing distance-to-pack and high score
ROC plot shows that distance-to-pack has slightly better discriminating power than high score. (Search type: search #1, described below.)

CALCULATING CONFIDENCE

Given that distance-to-pack has been found to be the metric with the optimal discriminating power, we can use an annotated data set to empirically calculate percent confidence for distance-to-pack as follows:

For each distance-to-pack d, percent confidence $= \frac{A(d)}{A(d)+Z(d)} \times 100\%$

where

$A(d)$ = number of correct peptide identifications in the annotated data set for distance to pack d
$Z(d)$ = number of incorrect peptide identifications in the annotated data set for distance to pack d

We perform a variety of database searches in order to study the robustness of our results under varying conditions:

Search #1: Non-zone modification search; MS/MS tolerance = 0.05 Da
Search #2: Non-zone modification search; MS/MS tolerance = 1 Da
Search #3: Zone modification search ±1000 Da; MS/MS tolerance = 0.05 Da
Search #4: Zone modification search ±1000 Da; MS/MS tolerance = 0.05 Da; modify Interrogator algorithm so that all MS/MS peaks are weighted equally

Note that this is a fairly severe test of robustness as searches #1/#2, #3, and #4 represent *three different algorithms.*

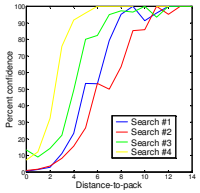

Figure 3. Percent confidence as a function of distance-to-pack for different database searches
The figure shows that percent confidence as a function of distance-to-pack varies significantly under different search conditions.

This lack of robustness to varying database search types makes it difficult to directly use distance-to-pack as a predictor for percent confidence. We have tried two different methods to try to address this problem. The first method involves using the average distance-to-pack of the data set to calibrate percent confidence. The second method is discussed below.

For each MS/MS spectrum let $F(s)$ represent the score distribution (histogram) – that is, $F(s)$ is the number of times each score s is found.

Calculate $G(s)$ as the cumulative score distribution: $G(s) = \sum_{s'=s}^{\infty} F(s')$

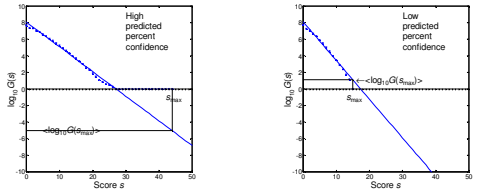

Figure 4. Cumulative score distribution $G(s)$ is used to predict percent confidence
Shown here are examples of the cumulative score distribution $G(s)$. For each MS/MS spectrum, we would like to estimate the probability that the peptide corresponding to the high score s_{max} is a true "hit" (match) rather than a random "hit." We hypothesize that the bulk of the distribution $G(s)$ arises from matches against random peptides and that the farther away s_{max} is from the bulk distribution, the higher the probability that the peptide corresponding to s_{max} is indeed a true hit. It is found empirically that, except at the tail near s_{max}, the decay of $\log_{10} G(s)$ as the score s increases is approximately linear. Thus, we model the bulk of the distribution $\log_{10} G(s)$ by performing a linear regression. The resulting linear least-squares model is then extrapolated and used to estimate $<\log_{10} G(s_{max})>$, the expected value of $\log_{10} G(s)$ at s_{max} due to random hits. According to the Poisson distribution, the probability of finding at least one random hit with score $s >= s_{max}$ is $1 - \exp(-10^{\wedge} <\log_{10} G(s_{max})>)$. Therefore, the statistical significance of a high score of s_{max} is $1 - [1 - \exp(-10^{\wedge} <\log_{10} G(s_{max})>)] = \exp(-10^{\wedge} <\log_{10} G(s_{max})>)$, and the predicted percent confidence based on the separation between s_{max} and the bulk of the distribution $G(s)$ is:

$$\text{Predicted percent confidence} = e^{-10^{-\log_{10} G(s_{max})>}} \times 100\%$$

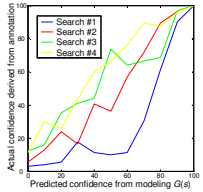

Figure 5. Comparison of predicted percent confidence and actual percent confidence derived from the annotated data set
Note that the predicted percent confidence calculated using the cumulative score distribution $G(s)$ does *not* rely on data set annotation. Thus, we can use the annotation to perform an *unbiased* test of the accuracy of the predicted percent confidence. This graph shows that the predicted percent confidence is reasonably accurate for a variety of MS/MS database search types. In addition, the discriminating power of this predicted confidence based on modeling $G(s)$ is comparable to distance-to-pack (data not shown).

CONCLUSIONS

We have presented some techniques for assessing the validity of peptide identifications made by a database search algorithm. All of our techniques require the presence of an annotated data set containing many MS/MS spectra since the annotation is used in the data analysis as well as in evaluating the effectiveness of various methods of data analysis. We find that for the Interrogator™ algorithm for database searching, modeling the cumulative score distribution $G(s)$ provides an estimated percent confidence that is reasonably accurate, has good discriminating power, and is robust under varying search conditions.

REFERENCES

1. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383.
2. Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75*, 435.
3. Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768.
4. Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, John Wiley & Sons, New York, 2001.
5. Swets, J. *A Science* **1988**, *240*, 1285.
6. Ott, L. *An Introduction to Statistical Methods and Data Analysis*, Duxbury Press, Boston, 1984.
7. Seymour, S. L.; Phu, L. M.; Tang, W. H.; Patel, A. A.; Lobodа, A.; Shilov, I. V.; Hunter, C. L.; Nuwaysir, L. M.; Settineri, T. A.; Schaeffer, D. A., 51st ASMS Conference on Mass Spectrometry (2003), Montreal, Canada, Poster Number: MPB 017.

TRADEMARKS/LICENSING