

Overview

Purpose: Study the validation of database search results of ETD spectra using ZCore, an MS/MS search algorithm.

Methods: The study looks into several schemes for generating false discovery rates of peptide identifications from tandem mass spectra and combination of forward and reversed protein sequences and compares the result with the expectation value calculations in ZCore. In ZCore we use a conditional expectation value, where we determine the expectation value of the best match under the condition that the second best match is an observed event.

Results: We show that the conditional expectation values determined in ZCore are highly specific, non-monotonic in the probability scores.

Introduction

Peptide/protein identification by tandem mass spectrometry and database search is a powerful, high throughput analytical tool. The applications of electron transfer dissociation (ETD) – a mechanism to fragment intact peptides by electron transfer, has widened the scope and the complexity of samples that can be analyzed with mass spectrometers¹.

For peptide identifications by tandem mass spectrometry and database search a major problem is validation of search results. A number of factors lead to incorrect peptide identifications. Some of these factors are unexpected modifications, the mass accuracy and incomplete peptide fragmentations. Many search engines generate a probability and expectation value of a match being a random hit. However, often these numbers are unreliable and researchers use alternative methods such as decoy database search to generate error rates. We compare two approaches to validations. Probability scores and expectation values indicate how unique the peptide match is with respect to the other candidate sequences from the database. The expectation values are corrected to account for multiple testing, for example by Bonferroni inequality. The power of this approach in practice has been that scoring is conservative, however, sensitivity could be improved.

The second approach uses decoy databases searches. The score distributions from the decoy databases form the Null Hypothesis. Then this distribution and scores from the real database are used to determine FDRs. Since proteomics tools are high throughput categorizations with FDR are attractive for analysis and interpretations. The disadvantage is that the same spectral data set is searched twice. The decoy database search approach does not use any additional features compared to the expectation value calculations from a regular database. It simply defines distribution of the false positives with respect to the best scores. It is also not obvious how large database should be to obtain adequate error rates. Large databases are expected to produce conservative error rates, possibly at the expense of sensitivity.

The use of decoy database for determining error rates in peptide identification significantly improves validation of database search results². It has been demonstrated that on average the likelihoods of a random match to a decoy and forward databases are similar.

FIGURE 1. Model Distributions

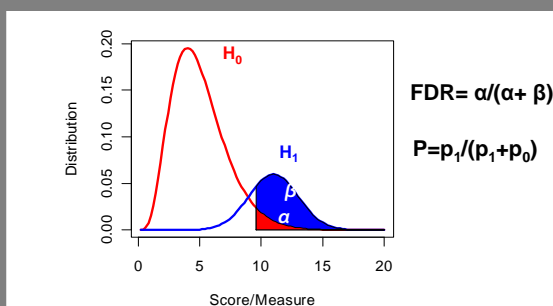


Table 1. Test of m Null Hypothesis (adopted from³).

$$FDR = E(Q) = \frac{V}{V+S} = E\left(\frac{V}{R}\right) \quad FPR = \frac{V}{m_0}$$

Statistical trials	Non-significant	Significant	Total
True Null Hypotheses (H ₀) (False hits)	U, True Negatives	V, False Positives	m ₀
Alternative Hypotheses (H ₁) (True hits)	T, False Negatives	S, True Positives	m-m ₀
Total	m-R=U+T	R=V+S	m

Methods

ZCore uses a compound probability model to identify sequence matches to a spectrum. The model employs shared peak counts and shared peak intensities to determine the probability that a match is a random event. The match, which is the least random event is the best match to the spectrum. Expectation values are normally determined from the Type I error – probability that the Null Hypothesis (the match is a random event) is true but falsely rejected:

$$E = N * (1 - (1 - P(x > X_{extreme}))^N)$$

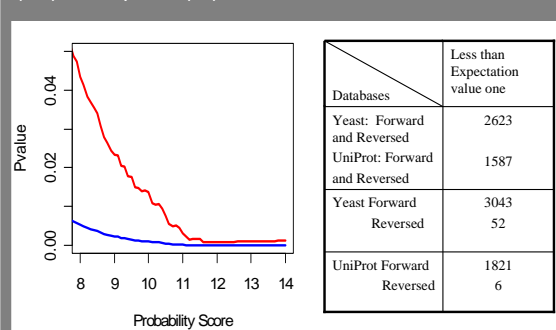
where X is the score of the best match, N is the number of candidate sequences. An approach to account for multiple testing by Bonferroni inequality divides the desired significance level, α , by the number of matches, S:

$$\alpha / S = E$$

The results from the Bonferroni correction tend to be conservative as many true positives are rejected. In the ZCore algorithm we use a conditional distribution to determine expectation values. In addition to the best score we use the distribution of the second best score as well:

$$E(X_1 | X_2) = N * P(X_1, X_2) / P(X_2)$$

FIGURE 2. P-values from single test (blue) and multiple tests (red).



In this approach, the expectation for the best match is determined from the assumption that the second best match is an observed event.

We compare these expectation values with the false discovery rates (FDR) from decoy database searches from large scale database searches. Benjamini and Hochberg³ developed a framework for computing FDR's for high throughput data. In this approach all p-values whose significance level is satisfied are sorted from lowest to the highest.

$$P_i \leq \frac{i}{m} q$$

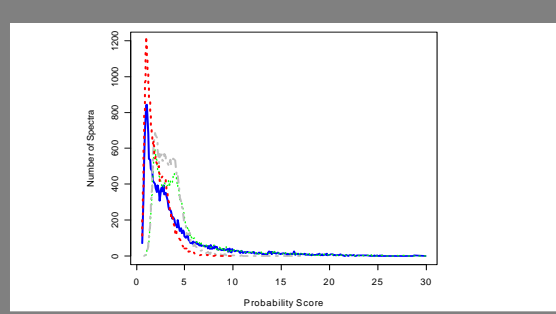
Figure 1 shows the model distribution of the Null and alternative hypothesis. The relevant quantities are defined in Table 1.

The data set used here is a spectral data set obtained from a fraction of whole cell yeast lysate. The spectra has been extracted by extract_msn and precursor charge states have been determined with Charger⁴ algorithm.

Results

We first demonstrate the effect of multiple testing on p-values. Figure 2 we show the p-values without (blue) and with (red) multiple testing effects for our data set. For smaller scores the single test p-values significantly underestimate the error rates. These results are obtained from separate searches of the dataset against reversed and forward databases. One of the features of this approach is that it is database size dependent.

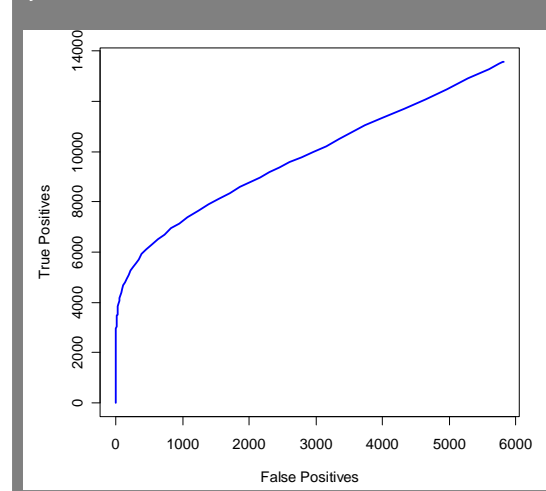
FIGURE 3. Distributions of ZCore probability scores against different databases, red (reversed yeast), blue (forward yeast), grey (reversed uniprot-sprot), green (forward uniprot-sprot).



We used a combination of different databases and search strategies to compute the FDR. The overall distributions of all spectra are shown in the Figure 3. First we searched the spectra using the yeast database, about 6400 entries. As it is expected the average database scores are larger for uniprot-sprot database (more than 300,000 entries) than for the yeast database. An interesting feature is that once the scores from the forward databases cross for the second time, they stay essentially equal. This is in contrast to the scores from the reversed databases. A similar behavior would be observed if the same peptides are identified in both searches.

In Figure 4 we show the ROC curves obtained from concatenated forward and reversed yeast and uniprot-sprot databases. As it is expected for a probability based identification model, the tail end of the curve is linear – probability of random matches to forward and reversed sequences are equal. At 1% FDR the numbers of true and false positives (as determined by matches to the forward and reversed sequences) are 3471 and 18, respectively. The first match to the reversed database is observed after 1681 matches to the forward sequences. Only 3 random matches observed amongst the 2974 matches to the forward database. When the non-unit expectation values are used (E<1), there are 2623 peptides from this database. There are 5 matches to the reversed sequences amongst these peptides. Only one of these spectra are charge state other than +2. The FDR of this result is about 0.04%. Thus 75% of all identifications with 1% FDR can be determined from expectation values. The result is consistent with those of other data sets. Thus for small database size, the yeast proteome has about 6500 proteins, concatenated database search identifications are both more sensitive and specific.

FIGURE 4. ROC curve obtained from the concatenated forward and reversed yeast databases.



There are some differences in results for the conditional expectation values and those computed from the reversed databases. As the database size increases, the identifications with non-unit expectation values decreases. The number of significant identifications is higher for separate forward and reversed database searches for both small (yeast) and large (uniprot-sprot) databases than for concatenated databases. This is in contrast with the FDR results from reversed databases where the FDR rates are always smaller for concatenated databases than for separate database searches. Another important feature of the conditional distribution is that it is non-monotonic in the first score. As it is seen from the Figures 3 and 4 the FDR's are almost-monotonic in probability scores of best matches. Since the conditional expectation depends on the first and the second scores it is not monotonic in the probability scores. Thus, it is quite often a case that first two best scores are high, Figure 5.

FIGURE 5. Distributions of second best match scores from forward database (blue), best match scores from the reversed database (red), and the false positives scores from the concatenated database (green).

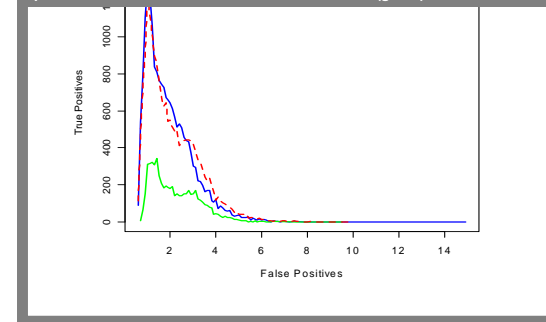
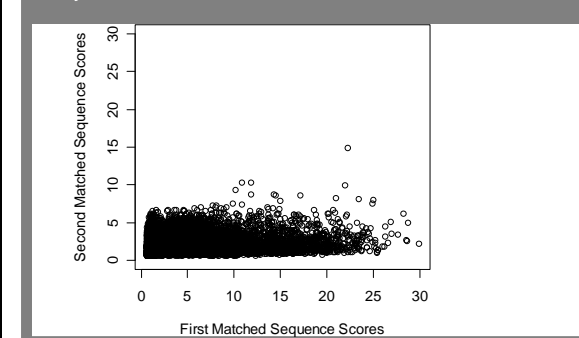


FIGURE 6. Scatter plot of the first and second best matched sequences' scores from yeast database.



The expectation is calculated from the assumption that the second best match is an observed event. Therefore the significance of the best score is measured with respect to the second match. In the case when the latter is a high scoring sequence, the significance of the first match is reduced. However, if the second best matches are not a high scoring sequence then even relatively low scores for the first matched peptides lead to significant matches, Figure 6.

Conclusions

We compare validations of database search results from ZCore using FDR calculations from reversed databases and expectation values of conditional distributions. Our study points out that while FDR calculations use only the scores of best matches, the conditional probabilities use first and second best matched scores for determining error rates. This approach is different from a more prevalent expectation value calculations where absolute value of the identification significance is used.

In practical terms, our approach allows to determine a large number (75%) of true positives with a very small FDR rate, less than 1%. This could be used as a quick test to assess general quality of the dataset. However, overall sensitivity of the results obtained with forward and reversed database searches using the best score matches is higher. Another important aspect of the conditional probability is that it is non-monotonic on the first score. As a result they can be used for additional identifications to improve the overall sensitivity.

References

- (1) Syka, J.E.; Coon, J.J.; Schroeder, M.J.; Shabanowitz, J.; D. F. Hunt, Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, *Proc. Natl. Acad. Sci., U.S.A.* 2004, v. 101, p. 9528-9533.
- (2) Elias, J.E.; Gygi, S.P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 2007, v.4, p.207-214.
- (3) Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of Royal Statistical Society, Ser. B.*, 1995, v. 57, p. 289-300.
- (4) Sadygov, R.G.; Hao Z.; Huhmer, A. Charger: Combination of Signal Processing and Statistical Learning Algorithms for Precursor Charge-State Determination from Electron-Transfer Dissociation Spectra, *Anal. Chem.* 2007, v. 80, p. 376-386.