Validating Database Search Results of ETD Spectra

Rovshan Sadygov, Gabe Zingaretti, Jim Shofstahl; Thermo Fisher Scientific, San Jose, CA, USA

Overview

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

Methods: We study four different schemes for generating false discovery rate (FDR) values. The schemes include: (1) a combination of both forward and reversed sequences and compute the expectation values and compare it, (2) compute a combination of forward and reversed sequences and compute the expectation values and compare it, (3) use decoy database searches, (4) compute an expectation value for each protein and compare the expectation value with the peptide FDR. In ZCore we use a combination 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 observe that the expected expectation values determined in ZCore are highly specific, sensitive, incorrect matches and putatively charge states have been determined with confidence.

Conclusions: We compare validations of database search results from ZCore using FDR calculations from reversed databases and expectation values of conditional distributions. Our data points out why FDR calculation use only the scores of best matches, the conditional probabilities calculate false positive matches, and sensitivity of the results obtained with multiple testing, where appropriate and the identification significance is validated.

Table 1. Test of Null Hypothesis (adopted from (2)).

<table>
<thead>
<tr>
<th>Test of Null Hypothesis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Hypothesis (TH)</td>
<td>0</td>
</tr>
<tr>
<td>True Negative (TN)</td>
<td>1</td>
</tr>
<tr>
<td>False Negative (FN)</td>
<td>m</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>m0</td>
</tr>
</tbody>
</table>

Figure 1 shows the model distribution of the null and alternative hypotheses. The relevant quantities are labeled in Table 2.

Methods

ZCore uses a combination probability model to identify sequence matches to a spectrum. The model consists of a peak score and a peak intensity to determine the probability that a peak is a sequence event. The peak, which is the most random event is the base peak intensity, which is the overall probability of a match.

Results

We first explore the effect of multiple testing on p-values. In Figure 2 we see the p-values where (blue) and (red) multiple testing effects for our data set. For another score the height and a significant p-value indicate the overall score. These results are consistent with other data sets obtained from yeast and uniprot-sprot databases. One of the features of the approach is the identification of true matches.

Conclusions

We compare validations of database search results from ZCore using FDR calculations from reversed databases and expectation values of conditional distributions. Our data points out why FDR calculation use only the scores of best matches, the conditional probabilities calculate false positive matches, and sensitivity of the results obtained with multiple testing, where appropriate and the identification significance is validated.