

Data and text mining

A predictive model for identifying proteins by a single peptide match

Roger Higdon1 and Eugene Kolker1,2,*

1The BIATECH Institute, Bothell, WA 98011, USA and 2Division of Biomedical and Health Informatics, University of Washington, Seattle, WA 98195, USA

Received on October 11, 2006; revised on November 17, 2006; accepted on November 20, 2006
Advance Access publication November 22, 2006
Associate Editor: Alfonso Valencia

Supplementary information: Supplementary Data are available at Bioinformatics online.

ABSTRACT

Motivation: Tandem mass-spectrometry of trypsin digests, followed by database searching, is one of the most popular approaches in high-throughput proteomics studies. Peptides are considered identified if they pass certain scoring thresholds. To avoid false positive protein identification, ≥2 unique peptides identified within a single protein are generally recommended. Still, in a typical high-throughput experiment, hundreds of proteins are identified only by a single peptide. We introduce here a method for distinguishing between true and false identifications among single-hit proteins. The approach is based on randomized database searching and usage of logistic regression models with cross-validation. This approach is implemented to analyze three bacterial samples enabling recovery 68–98% of the correct single-hit proteins with an error rate of <2%. This results in a 22–65% increase in number of identified proteins. Identifying true single-hit proteins will lead to discovering many crucial regulators, biomarkers and other low abundance proteins.

Contact: ekolker@biatech.org

1 INTRODUCTION

High-throughput studies to characterize cellular protein contents most often use tandem mass spectrometry (MS/MS) of trypsin digests of the protein sample followed by database searching of protein sequences ( Aebersold and Mann, 2003; Kolker et al., 2006). Proteins are identified based on matches to their sequences resulting from comparing of observed peptide MS/MS spectra to theoretical spectra. The latter being generated from possible peptides in the protein sequence database. Peptides are considered identified if they pass certain preset scoring thresholds based on the native scores of the database search algorithms ( Pang et al., 2002; Perkins et al., 1999; Washburn et al., 2001) or after applying probabilistic models (Higdon et al., 2004; Keller et al., 2002; Nesvizhskii et al., 2003). Though these thresholds are aimed at reducing the false positive error rate, still, a significant number of false peptide identifications can occur in high-throughput analyses due to the immense volume of the spectra (Cargile et al., 2004; Kolker et al., 2006). Often hundreds of thousands spectra are being produced per typical experiment, so even a very low error rate will result in many false peptide identifications (Kolker et al., 2006). In turn, these false peptide assignments can result in many false positive protein identifications (Kolker et al., 2006; Nesvizhskii et al., 2003). Therefore, more conservative criteria have been recently (Carr et al., 2004; Omenn et al., 2005) recommended requiring ≥2 unique peptides to be identified within a single protein for its positive identification (double-hit proteins). Currently, there is heated debate in scientific community as to whether such criteria should be made part of publication and reporting standards for proteomics studies (Bradshaw et al., 2006; Carr et al., 2004; Orchard and Ping, 2006).

The problem with the two-peptides-per-protein approach is that in a typical experiment hundreds of proteins are identified only by a single peptide match (single-hit proteins). This may be due to many factors such as low concentration, few tryptic peptides in small proteins, or masking by other more expressed proteins. Simply ignoring these proteins will result in the loss of unique information, hindering our path to a deeper understanding of cellular protein contents, finer regulatory processes, and next-generation biomarkers.

Randomized sequence databases have been routinely used in bioinformatics research for over 30 years (Doolittle, 1986) and more recently used in proteomics to estimate the number of incorrect or random peptide matches produced by the database search (Pang et al., 2002; Qian et al., 2005). The underlying assumption for a usage of the randomized database is that the probability of a peptide match to a randomized sequence is equal to the probability of an incorrect match to the original sequence database. This assumption holds in a simultaneous search of a combined database composed of the organism’s original protein sequences and their randomized versions since it avoids bias created by true peptide spectra matching erroneously to randomized sequences when the databases are searched separately (Beausoleil et al., 2006; Elias et al., 2005; Higdon et al., 2005). Hence, discriminating between correct and incorrect peptide identifications (or single-hit protein identifications) is equivalent to discriminating between original and reshuffled sequence matches.

Here, we introduce a method for distinguishing between true and false identifications among single-hit proteins. The first step is based on a use of randomized database search to estimate the rate of false matches. Next, logistic regression models (McCullagh and Nelder, 1999) identify the best predictors and assign probabilities for single-hit proteins. Finally, cross-validation is employed to remove possible biases in the estimation of error rates. The entire approach is implemented in whole proteome studies of three bacterial samples. The first two samples analyzed are Rhodobacter sphaeroides wild-type and mutant strains, and the third is a Shewanella...
oneidensis sample. The samples were subjected to multiple MS/MS runs on LCQ and LTQ ion trap mass spectrometers. The acquired peptide spectra were searched by SEQUEST (Eng et al., 1994), and peptide identification probabilities were generated and re-adjusted with our logistic identification of peptide sequences (LIPS) model (Higdon et al., 2004) (see Methods). As a result, we were able to recover 68–98% of the correct single-hit proteins (with <2% error rate) that led to a 22–65% increase in the number of identified proteins.

2 METHODS

2.1 Samples and MS/MS analysis

The wild-type and mutant strains of R.spheroides were grown in batch cultures under anaerobic condition with dimethyl sulfoxide (DMSO) by the Kaplan lab at the University of Texas in Houston (Tai and Kaplan, 1985). S.oneidensis cells were grown in chemostates under suboxic conditions with fumarate at the Pacific Northwest National Laboratory (Elias et al., 2005; Kolker et al., 2005). Two separate whole cell lysates (two biological replicates) of each the wild-type and mutant R.spheroides samples were digested with trypsin and subjected to eight replicate liquid chromatography (LC) MS/MS spectrometry. LC-MS/MS runs using a 240-min LC gradient on an LCQ Deca XP PLUS (Thermo Electron) ion-trap mass spectrometer. Similarly, two biological replicates of S.oneidensis cells were digested with trypsin and each subjected to three 90-min gradient LC-MS/MS runs on an LTQ (Thermo Electron) ion-trap mass spectrometer (Elias et al., 2005; Kolker et al., 2005).

Each MS/MS run was searched with SEQUEST (Eng et al., 1994) using 2.5 Dalton parent mass accuracy and an unconstrained enzyme search at charge states 1 through 3 for each acquired spectrum. Three experimental data sets (two for R.spheroides and one for S.oneidensis) were searched against combined databases containing the original protein sequences of the corresponding organism and reshuffled versions of these sequences appended together (Higdon et al., 2005). Reshuffled (randomly permuted) sequences were used rather than reversed sequences because they were found to perform slightly better in our previous work (Higdon et al., 2005).

2.2 LIPS model

The SEQUEST search output files were processed with our LIPS model to provide peptide identification probabilities based on a logistic regression model (Higdon et al., 2004). LIPS uses a linear combination predictors that currently include cross-correlation score ($X_{corr}$), relative difference between the first and second highest cross-correlation scores ($\Delta C_r$), peptide length, charge state, the number of tryptic termini, and whether the rank preliminary score is <5.

An adjustment to the LIPS probability is done based on the basic randomized or reshuffled database assumption: probability of reshuffled identification ($P_{RS}$) equal to probability of incorrect identification from an original sequence. This assumption has been validated in our previous work (Higdon et al., 2005) and that done by others (Beausoleil et al., 2006; Elias et al., 2005). We further demonstrate the validity of this assumption in Figure 1, where the distribution of the peptide scores from our LIPS are shown to be nearly identical between reshuffled sequences matches and incorrect original sequence matches from MS/MS runs of our previously published standard protein mixture (Purvine et al., 2004). We estimate $P_{RS}$ by fitting a polynomial logistic regression model to the linear predictor generated by the LIPS model. The probability of a correct peptide match (given it is not a reshuffled sequence) is generated by a simple application of Bayes’ Rule and the basic randomized database assumption resulting in the estimate of $P_{Correct}$ as follows:

$$P_{Correct} = \frac{1 - 2 \cdot P_{RS}}{1 - P_{RS}} P_{RS} < 0.5 \quad (1)$$

![Fig. 1. Distribution of LIPS index values comparing incorrect original sequence matches to reshuffled matches shows the distribution of LIPS scores generated from a MS/MS data set of two MS/MS runs from a 12 protein mixture run on an LTQ using an identical protocol to that used for the MS/MS runs of S.oneidensis described in the Section Methods. There were a total of 4955 (49%) incorrect original sequence matches and 5212 (51%) reshuffled sequence matches. The distributions of incorrect original-sequence matches and reshuffled-sequence matches appear to be nearly identical.](image)

Table 1. Summary of the peptide and protein identification

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Unique peptide IDs</th>
<th>Proteins 2+ peptides</th>
<th>Single-hit proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original RS</td>
<td>Original RS</td>
<td>Original RS</td>
</tr>
<tr>
<td>R.spheroides WT</td>
<td>3096</td>
<td>204</td>
<td>454</td>
</tr>
<tr>
<td>R.spheroides MU</td>
<td>3031</td>
<td>196</td>
<td>450</td>
</tr>
<tr>
<td>S.oneidensis</td>
<td>3969</td>
<td>68</td>
<td>661</td>
</tr>
</tbody>
</table>

Spectra identified using an adjusted LIPS probability >0.9. The estimate of the number of correct is given by subtracting the number of reshuffled (RS) sequences from the number of original sequences.

An adjusted LIPS probability threshold of 0.9 was used to accept a peptide assignment as ‘correct’ and summarized in Table 1.

2.3 Statistical analysis

Data sets were created for each experiment containing all proteins identified by a single unique peptide for the original and reshuffled sequences. A number of predictor variables for each protein were considered for the purpose of discriminating between correct and reshuffled protein sequences. These predictors included: the total number of identified spectra, the protein length, the percentage of the protein sequence covered by the identified peptide, whether multiple forms of the peptide were identified (fragments or other charge states), whether the peptide was identified in multiple samples, whether the peptide was identified in multiple MS/MS runs of a single sample, the sum of adjusted LIPS probabilities, the maximum adjusted LIPS probability across all identified spectra, the peptide length, and the number of tryptic cleavage points. For each predictor the number of the original and reshuffled sequences meeting a given threshold and the odds ratio of original versus reshuffled was calculated. Logistic regression models of the probability of being reshuffled were fit individually and jointly to the
Table 2. Frequencies of the protein identification for original and reshuffled (RS) sequences based on probabilities estimated by leave-one-out cross-validation and logistic regression including the maximum LIPS probability and the peptide length at different thresholds

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$P^* &gt; 0.9$</th>
<th>$0.75 &lt; P &lt; 0.9$</th>
<th>$P &lt; 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original RS</td>
<td>% correct (95% LB)$^b$</td>
<td>% Rec.$^c$</td>
</tr>
<tr>
<td><em>R.spheroides</em> WT</td>
<td>103</td>
<td>2</td>
<td>98.1 (91.2)</td>
</tr>
<tr>
<td><em>R.spheroides</em> MU</td>
<td>112</td>
<td>1</td>
<td>99.1 (93.8)</td>
</tr>
<tr>
<td><em>S.oneidensis</em></td>
<td>420</td>
<td>7</td>
<td>98.3 (96.2)</td>
</tr>
</tbody>
</table>

$^aP$ corresponds to the probability derived from logistic regression model and use Equation (1).

$^b$95% lower-bound (LB) based on Poisson model

$^c$Percentage of single-hit proteins recovered at the 0.9 threshold, calculated by the number of original minus reshuffled sequences divided by the number of total original minus reshuffled sequences (shown in Table 1).

$^d$95% upper-bound (UB) based on Poisson model.

3 RESULTS AND DISCUSSION

The total number of double- and single-hit proteins (both original and reshuffled sequences) for each experiment is summarized in Table 1. In the *R.spheroides* wild-type experiment, there were 334 single-hit proteins from the original sequence database versus 188 from reshuffled sequences. Assuming that 188 of 334 proteins are incorrect, the conservative two-hits-per-protein approach is estimated to miss 146 ‘correct’ protein identifications. The results are virtually identical for *R.spheroides* mutant samples. The proportion of reshuffled hits is much lower in *S.oneidensis* samples due to the greater amount of high quality and high scoring peptides (Table 1).

Supplementary Table 1 summarizes the performance of the different predictors (described in Section Methods) at discriminating between correct and reshuffled single-hit proteins. Individual predictors that discriminate well are those that are (1) statistically significant in the single variable logistic models; (2) have large odds ratios; and (3) have large numbers of proteins exceeding the given thresholds. Upon adjusting for all the other predictors, the maximum LIPS probability (across all matched spectra) and the peptide length are the only highly significant predictors (Supplementary Table 1, $P$-values <0.0001 and 0.0006, respectively). After adjusting for these two predictors the remaining ones are no longer statistically significant. (This is likely due to the inter-relatedness of these predictors, for example, percent coverage is a combination of the peptide and protein lengths.) Very similar results were obtained for the *R.spheroides* mutant samples (Supplementary Table 1). Because of the low number of reshuffled protein matches in the *S.oneidensis* experiment, the potential for over-fitting these models increased. Nevertheless, here again very similar results to the *R.spheroides* samples were observed with only the peptide length and the maximum LIPS probability remaining significant (Supplementary Table 1).

The model containing the best two predictors was fit to the data and Equation (1) (see Methods) was used to assign probabilities to the single-hit proteins. The fitted probabilities were divided into three categories; those >0.9 were considered to very likely to be correct and <0.75 there were near equal numbers of original and reshuffled sequences indicating a low likelihood of being correct. In the *R.spheroides* wild-type experiment a probability threshold of 0.9 results in an estimated 98.1% (95% lower-bound of 91.3%, based on cross-validation estimates) of correct protein identifications out of 103 total (Table 2 and Section Methods). A probability between 0.75 and 0.9 results in a lower estimate of 80.4% (95% lower-bound of 60.3%) and with a probability <0.75 very few proteins (4.8%) are estimated to be correct (95% upper-bound of 21.5%). In the *R.spheroides* wild-type experiment estimate of 101 (103-2) correct original sequences exceeding the 0.9 threshold represent 69% (101 out of 146) of the total estimated correct single-hit proteins (Supplementary Table 2 online). Very similar error rates to the *R.spheroides* wild-type were obtained for the mutant and *S.oneidensis* samples, except the in the latter case a much greater proportion of the proteins exceeded the 0.9 threshold. The percentages of 68% (111 out of 164) and 98% (413 out of 421) for the *R.spheroides* mutant and *S.oneidensis* samples, respectively, indicate that a majority of the single-hit protein identifications can be recovered (Supplementary Tables 2 and 3). Most of the remaining correct single-hit proteins can be recovered as well, if one is willing to accept the higher error rate associated with using the 0.75 probability threshold.

While we have demonstrated that all single-hit proteins are not equal, it is also clear that double-hit proteins are not equal either. For example, eight double-hit protein identifications in the *R.spheroides* wild-type sample corresponded to reshuffled sequences versus 105 original sequences that exactly double-hit (Table 1). This suggests the necessity for further evaluation of even double-hit protein identifications. The small numbers of such hits to reshuffled sequences make this situation difficult to model. Therefore, we applied our single-hit model to this data using the maximum LIPS probability (over all identified spectra) and the maximum length of identified peptides as
predictors. The model worked extremely well at distinguishing between the original and resheluffle sequences. There were 81 double-hit proteins matched to original sequences >0.9 probability threshold versus none to resheluffle sequences; an estimated 100% correct identifications (95% lower-bound of 93.8%). There were seven matches each to original and reshuffled sequences <0.75 threshold; an estimated 0% correct (95% upper-bound of 56.3%). Virtually identical results were obtained for the R.spheroïdes mutant sample and the single reshuffled double-hit protein in the S.oneidensis sample had an estimated probability of only 0.3, while 208 out the 209 exactly double-hit original proteins had probabilities greater >0.9. As the number of MS/MS runs and peptide spectra increase, so does the likelihood of two random matches to the same protein. While explicitlily modeling these situations is likely to prove difficult given the small numbers of false-positive identifications, our single-hit approach proved to be extremely effective in this situation as well.

4 CONCLUSION

The obtained results demonstrate that not all single-hit proteins were created equal. A simple model using the peptide length and the maximum LIPS probability was the best discriminatory model. This approach recovered from 68 to 98% of the potential correct single-hit protein identifications. A threshold of 0.9 for the estimated protein probability resulted in a 22–65% increase of the identified proteins with the error rate of <2%. Therefore, confident and substantial high-throughput protein identification can be made based on the single-hit proteins. This is especially vital in regards to the many biologically important low abundance proteins that are currently often disregarded, as they are identified only by a single peptide. With the introduced models individual protein identification can be assessed better and the number identified proteins can be increased significantly. In turn, this will lead to a deeper understanding of cellular protein contents, discovering finer regulatory processes, and finding next-generation biomarkers.

ACKNOWLEDGEMENTS

We greatly appreciate the labs of Samuel Kaplan, Jim Fredrickson, and Richard Smith and Jason Hogan for their preparation of the samples. We also thank Gerald van Belle, Jason Hogan, and Natali Kolker for their fruitful discussions and comments. This work was supported by the DOE’s OBER grant DE-FG08-01ER63218 and the NIGMS R01 grant GM076680-01A1 to E.K.

Conflict of Interest: none declared.

REFERENCES