An Objective Organism-Based Evaluation of Tandem Mass Spectrometric Data Obtained from Proteomic Studies

Konstantinos Thalassinos, Georgios Efstathiou, Susan E. Slade, James H. Scrivens

Biological Mass Spectrometry and Proteomics, Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL, U.K.

OVERVIEW

Aims

The objective of this work was to develop a framework that allows the objective evaluation of MS/MS protein identifications, in particular one hit wonders, and to do so in an organism-specific manner.

Methods

Entire proteome sequences for organisms studied were obtained from the Integr8 web site in fasta format. A program digested each protein sequence *in-silico*. For each peptide all successive *n*mers (*n* = 2-8) were extracted and the uniqueness of each peptide based on varying ppm error values was calculated. A public shotgun proteomics dataset, was used to evaluate one hit wonder identifications. Spectra were *de-novo* sequenced with Peaks and a Java program extracted confident tags and searched these against our organism specific database.

Results

A detailed analysis of the publicly available shotgun dataset revealed that the majority of proteins (835 / 1574) identified were based on a single peptide. Detailed analysis of these identifications using our organism specific database revealed that most of the one hit wonders were false positive identifications. Finally, our newly developed framework revealed 7 new proteins not previously identified from that dataset.

INTRODUCTION

The rapid development of mass spectrometry-based proteomics has led to the generation of large amounts of experimental data. Database search programs use this data to identify the peptide/protein products. There are several such algorithms in general use, which share a common feature in that they always report peptide/protein identifications irrespective of the quality of the data used to perform the search. False positive identifications can arise from a number of possible causes, these include:

- Low quality of MS/MS spectrum e.g. poor signal to noise.
- The peptide sequence is not in the selected database.
- Post translational modification of the selected peptide.
- Algorithm specific search criteria. Submitting the data to more than one search engine may give different results.
- The presence of degenerate peptide structures in databases due to close homology or alternative splice forms.

As a rule of thumb proteins identified with two or more peptides are considered to be confidently identified. Proteins identified with one peptide, also referred to as one hit wonders, can lead to a large number of false positive identifications. We report an organism-specific study on the validity of including one hit wonders as confident identifications.

MATERIALS AND METHODS

Obtaining Proteome Sequences

Proteomes for all organisms apart from SPM2 (a phage infecting cyanobacteria) were obtained from the Integr8 web site (http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do). Each proteome was downloaded as a Fasta formatted file. The proteome of SPM2 was created by combining the outputs of gene prediction programs (Glimmer and GeneMark) and the Blast results of all open reading frames. The organisms analysed as part of this study are listed in **Table 1**.

Organism	Number of Proteins
Drosophila melanogaster	16256
Saccharomyces cerevisiae	5799
Escherichia coli K12	4329
Escherichia coli W3110	4179
Bacillus licheniformis Goettingen	4179
Synechococcus sp. WH8102	2512
SPM2	249

Table 1. Proteomes obtained from the Integr8 web site1 apart from SPM2 which was created in-house

Creation of an organism-specific database

The processing of each proteome was performed using a program written in ANSI C. The data generated from the program were saved in a MySQL (http://www.mysql.com/) database in a number of different tables. Each protein was *in-silico* digested using the cleavage specificity of trypsin. The sequence and the monoisotopic molecular weight for each peptide were saved in the database.

The program extracted all successive *n*mers, where *n* = 2 - 8, from each peptide sequence and saved these as well. A set of SQL scripts, which were automatically generated by the program, were used to calculate the frequency of each *n*mer. Another set of SQL scripts were used to calculate the uniqueness of a peptide mass based on varying ppm error values. The mass of a single peptide and the ppm error value were used to calculate a mass error range. For a mass of 1000 Da and a ppm error of 100 the mass error range would be the mass interval from 1000-0.1 = 999.9 to 1000 + 0.1 = 1000.1 Da. The mass error range was used to count the number of peptide masses that were present within that window. Unique peptide masses resulted in a count of 0,which meant that there were no other peptides within that given mass error range.

These values were also stored in the database. All the tables in the database hold non-redundant information. An overview of the approach is shown in **Figure 1**.

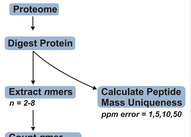

Figure 1. An overview of the *in-silico* analysis of each proteome.

Analysis of the Public Shotgun dataset

The public shotgun dataset [1] (accession opd00034_YEAST) was downloaded from http://bioinformatics.icmb.utexas.edu/OPD/. This included all the MS/MS data and SEQUEST output for each of these spectra. Protein identification information (protein accession and sequence, MS/MS datafile name, peptide sequence Xcorr values etc.) was included in an Excel spreadsheet. The data in the Excel file were exported to a tab delimited text file. A Perl script used this file to insert the information into a MySQL database. This database was queried to identify all proteins identified with only one peptide. The MS/MS data resulting in those peptide sequences were selected for further processing by Peaks.

***De-novo* Sequencing**

Peaks [2] (Bioinformatics Solutions Inc.) is a program that performs *de-novo* sequencing of peptide MS/MS spectra and provides derived complete, and partial peptide amino acid sequences together with confidence limits. The program is capable of operating in an automated fashion enabling large data sets to be processed efficiently. For the Prince *et al.*, dataset, obtained on an Ion Trap, a parent and fragment mass error tolerance of 1 Da was used. The instrument was set to Ion Trap and the selected enzyme to trypsin. Carbamidomethylation of cysteines was set as a fixed modification. Peaks data of the top 5 peptide sequences was exported in HTML format. The HTML file contained the datafile name, the *m/z* and charge of the precursor, the sequence of each peptide, the overall confidence for each peptide expressed as a percentage and the positional confidence for each amino acid. A Perl script was written to parse the Peaks HTML files and save this output in a MySQL database. In addition to the information contained within the HTML file, the script processed each peptide sequence so that the sequence and position of each confident tag on the peptide and any non-confident amino acids C-terminal to the tag were also saved.

Tag-mediated search

A program, written in Java 1.5.0_03, was used to search the confident tags from each precursor against the processed proteome of the corresponding organism. An overview of the search process is presented in **Figure 2**.

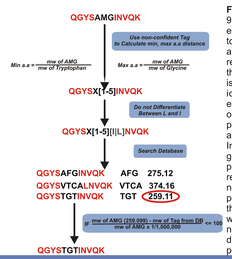

Figure 2. All confident tags (more than 90%)were combined to form a regular expression. The non-confident tags were used to calculate a minimum and maximum number of amino acids between confident tags. The regular expression was constructed so that there was no differentiation between leucine and isoleucine residues since these residues have identical molecular weight. The regular expression was used to extract all peptides from our database that matched these criteria. The peptides returned from this search were filtered as follows.

In a number of cases the regular expression generated was non-tryptic (contained a K or R prior to the C-terminus). In such cases using the regular expression to search the peptides would not have resulted in any identifications, since all peptides in our database were fully tryptic. For this reason, non-tryptic regular expressions were used to create fully tryptic ones. These newly tryptic peptides, in the majority of cases, did not contain the information necessary to filter peptides returned from the database.

RESULTS

Analysis of the Organism-Specific Database

Using our organism specific database we were able to calculate the tryptic peptide distribution (**Figure 3**), the length of an *n*mer required to achieve peptide uniqueness (**Figure 4**) and the effect of mass accuracy on peptide uniqueness (**Figure 5**).

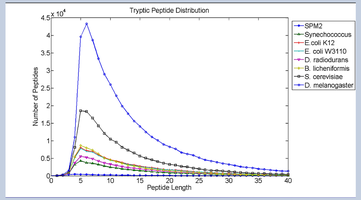

Figure 3. Tryptic peptide distribution of the organisms under study. For all organisms the majority of fully digested tryptic peptides have a length of 7 amino acids.

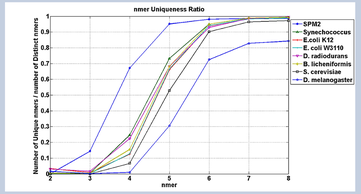

Figure 4. Plot of peptide uniqueness. The x axis describes the number of the nmer. The y axis describes the ratio of the number of unique *n*mers (*n*mers that exist only once in the database) divided by the total number of distinct *n*mers (the number of *n*mers with different amino acid sequence). A value of one indicates that all *n*mers present in the database are unique. Please note that for *n*mer length of 2 some organisms appear to have a small number of non-unique sequences. These are due to the database containing protein sequences with the X (any amino acid) character. These proteins were not excluded from our study but interrogation of our databases showed that there were very few such sequences and therefore they did not affect any subsequent analyses. The graph illustrates that for very small proteomes such as that of SPM2, uniqueness of sequences (at the 95% level) is achieved when selecting a 5mer while for organisms such as E. coli the same is achieved at 6mer length. For even larger organisms like *D. melanogaster* not even an 8mer can achieve the same uniqueness specificity.

Analysis of the publicly available dataset

Figures 6 and **7** show the number of MS/MS spectra collected and the database search results for the Prince *et al.*, dataset. More than half of the protein identifications were based on a single peptide. The MS/MS spectra giving rise to these identifications were selected and processed using our tag-mediated search. The results are summarised in **Figures 8-10**. **Figure 8** shows the number of MS/MS spectra filtered at each stage. **Figure 9** shows the number of proteins identified from the Prince *et al.*, dataset and **Figure 10** compares the results between our tag-mediated search and SEQUEST regarding one hit wonder protein identifications.

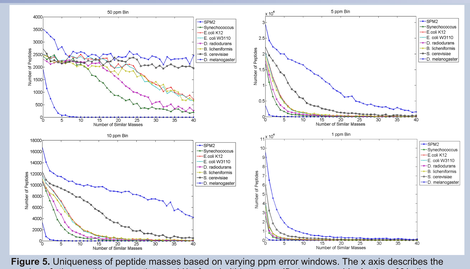

Figure 5. Uniqueness of peptide masses based on varying ppm error windows. The x axis describes the number of other peptide masses that would be found within the specified ppm error bin. A value of 0 indicates that a peptide mass is unique. For organisms with a small proteome using a ppm error of 1 allows identification of unique peptides while for larger proteomes even this kind of mass accuracy is not enough to uniquely identify a peptide.

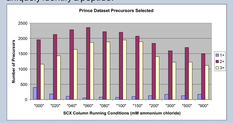

Figure 6. Statistics collected from the Prince *et al.*, Dataset.

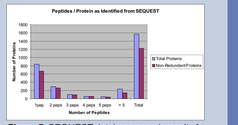

Figure 7. SEQUEST database search results for the Prince *et al.*, dataset.

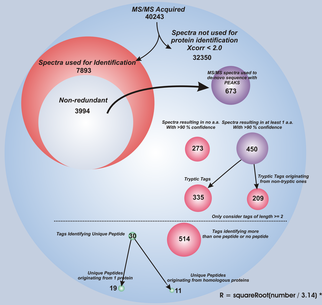

Figure 8. Number of MS/MS spectra filtered at each stage of the database search and tag-mediated search process.

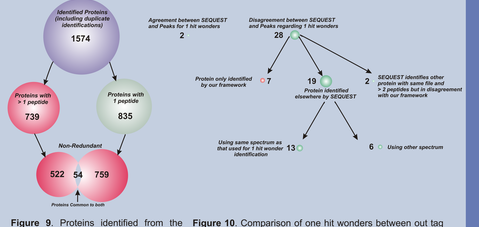

Figure 9. Proteins identified from the Prince *et al.*, dataset. The majority of protein identifications were based on a single peptide.

Figure 10. Comparison of one hit wonders between out tag mediated search and SEQUEST. In only 2 out of the 30 cases did our results agree with those from SEQUEST. Using our approach we were able to identify 7 additional proteins that were absent from the SEQUEST result set.

CONCLUSIONS

An oligomer-segmented organism specific database approach has been developed which when combined with de-novo sequencing information provides objective evaluation of the information content of peptide MS/MS spectra.

The effect of mass accuracy, tag length and tag sequence on peptide uniqueness was examined using our novel oligomer segmented, organism specific database.

For small proteomes mass accuracy can uniquely identify peptides while for larger ones it may not be sufficient.

The length of a sequence tag that achieves peptide uniqueness varies for each proteome but using our framework it can be accurately calculated.

One-hit-wonders should be carefully examined: Discarding them will lead to important low abundance proteins being overlooked while accepting all of them in the list of identified proteins will lead to erroneous conclusions regarding the biology of an organism.

Seven additional proteins were only identified by using our framework.

Our framework allows the critical evaluation of one hit wonders and does this on an organism specific basis.

REFERENCES

[1] Prince, J.T., Carlson, M.W., Wang, R., Lu, P., Marcotte, E.M. (2004) The need for a public proteomics repository. *Nature Biotechnology* 22: 471 - 472.

[2] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie., G. (2003) PEAKS: Powerful Software for Peptide De Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* 17: 2337-2342.

ACKNOWLEDGEMENTS

K. Thalassinos would like to thank the Welcome Trust for funding provided.