Published in final edited form as:

J Proteome Res. 2011 April 1; 10(4): 2088–2094. doi:10.1021/pr101157s.

FDRAnalysis: A tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines

David C Wedge¹, Ritesh Krishna², Paul Blackhurst², Jennifer A Siepen¹, Andrew R. Jones², and Simon J. Hubbard^{1,*}

¹Faculty of Life Sciences, University of Manchester, Manchester M13 9PT

²Department of Comparative Molecular Medicine, School of Veterinary Science, University of Liverpool, Liverpool L69 7ZJ, UK

Abstract

Confident identification of peptides via tandem mass spectrometry underpins modern high-throughput proteomics. This has motivated considerable recent interest in the post-processing of search engine results to increase confidence and calculate robust statistical measures, for example through the use of decoy databases to calculate false discovery rates (FDR). FDR-based analyses allow for multiple testing and can assign a single confidence value for both sets and individual peptide spectrum matches (PSMs). We recently developed an algorithm for combining the results from multiple search engines, integrating FDRs for sets of PSMs made by different search engine combinations. Here we describe a web-server, and a downloadable application, which makes this routinely available to the proteomics community. The web server offers a range of outputs including informative graphics to assess the confidence of the PSMs and any potential biases. The underlying pipeline provides a basic protein inference step, integrating PSMs into protein ambiguity groups where peptides can be matched to more than one protein. Importantly, we have also implemented full support for the mzIdentML data standard, recently released by the Proteomics Standards Initiative, providing users with the ability to convert native formats to mzIdentML files, which are available to download.

Keywords

bioinformatics; false discovery rate; multiple search engines; web server; data standards

Introduction

A major challenge of any tandem mass spectrometry experiment in the proteomics field is to differentiate correct from incorrect candidate peptide identifications used to infer protein identities. The scale of modern high-throughput proteomic analyses makes the manual verification of individual peptide-spectrum matches (PSMs) impractical, and automated methods have become increasingly popular. The use of decoy databases, containing reversed or randomised proteins, has proved to be a particularly fruitful method for estimating false

*To whom correspondence should be addressed. simon.hubbard@manchester.ac.uk, tel: +44 (0)161 306 8930, fax: +44 (0)161 275 5082 .

Supporting Information Available: File 1. Sample output file in tab-separated format containing a summary of the results (number of true positives, false negatives and calculated FDR for individual search engines and for the consensus approach) a list of peptides identified by FDRAnalysis and information for each peptide including: sequence, spectrum ID, parent protein, identifying search engine(s), FDR score, position within protein, calculated and experimental masses, charge and mods. This material is available free of charge via the Internet at <http://pubs.acs.org>.

discovery rates (FDR) ¹. An attractive property of the FDR approach is its ability to overcome the multiple testing problem, where many individual candidate PSMs are generated in a proteomic experiment with a variety of individual confidence scores. By selecting a global FDR one can select the set of PSMs that provides a user-specified overall error rate. Coupled to the conceptual ease of the general approach this has led to a large number of studies using this approach²⁻⁹, despite some recent controversies in its use¹⁰.

In addition to the variation in FDR estimation methodology, differences between the various database search engines themselves leads to different sets of candidate PSMs. However, the results from several engines can be combined to improve sensitivity. Indeed, we recently demonstrated that combining result sets can not only enhance the number of peptide identifications but also increase confidence in these identifications¹¹. In this earlier work, we introduced the concept of an *FDRScore*, which is derived by smoothing the distribution of q-values associated with the PSMs from individual search engines (a q-value is the minimum FDR threshold at which a given score would be called significant). A second measure, the *combined FDRScore* effectively up-weights the confidence in PSMs found by multiple search engines.

Here, we present a web-enabled implementation of these algorithms, FDRAnalysis, which enables the upload of peptide identification results from target/decoy searches carried out by three different search engines - Mascot ¹² and two open source tools, OMSSA ¹³ and X! Tandem ¹⁴. Our web tool supports both native file formats from the various search engines (*dat* for Mascot, *csv* for OMSSA and *xml* for X!Tandem) and, crucially, the new peptide/protein identifications XML standard, mzIdentML, recently released by the Human Proteome Organisation - Proteomics Standards Initiative (HUPO-PSI, <http://www.psidev.info/>). Importantly, FDRAnalysis can import native format search results, convert them to mzIdentML prior to analysis, or deal directly with the mzIdentML format. The internal processing has been re-factored to use mzIdentML at each stage. The combined results are produced as an mzIdentML file, as well as tab-separated values (TSV) and comma-separated values (CSV) files. Both the converted search engine outputs and the output of the FDR analysis are available to the user. FDRAnalysis can therefore act as a converter from selected 'native' to mzIdentML formats, which we believe will be of considerable use to the proteomics community.

In addition to the *combined FDRScore* calculations, we also present a series of graphical analyses which can be used to investigate the results of the target/decoy strategy and reliability of their data. This includes some common analyses which consider the relative numbers of matches to target and decoy databases, the position of peptide hits in the parent proteins, and mass defects of target and decoy hits.

Although the FDRAnalysis tool is valuable to the general proteomics research community regardless of experiment type, it was originally developed as part of an N-terminal peptide identification pipeline, in which the N-terminal peptides from a protein are isolated and analysed independently ¹⁵. The tool therefore has an option to analyse the peptide position in the protein, reporting the frequency of PSMs at position one or two in the protein sequence (allowing for initiator methionine cleavage). This information can prove useful in looking at the quality of N-terminal preparation, for which large numbers of internal peptides would suggest that either the method is 'leaky', i.e. it is not specifically selective for N-terminal peptides, or that there are many false positive identifications.

We believe FDRAnalysis will be useful to proteome scientists wishing to increase the sensitivity and confidence of candidate peptide identifications from multiple search engines, using a unified statistical treatment delivering FDR scores. Importantly, FDRAnalysis

outputs mzIdentML, converting search engine file formats into this HUPO-PSI supported format.

Experimental section

Algorithm implementation

The FDRAnalysis tool is based on our previously published algorithm which performs an integrated FDR analysis over multiple search engines¹¹, and is made available via the following URL: http://www.ispider.manchester.ac.uk/FDRAnalysis/FDR_analysis_home.html.

In brief, the suite is able to carry out the following tasks (Figure 1),

- *Convert native formats to mzIdentML.* We have developed in-house parsers in Perl for converting OMSSA CSV and Tandem XML results to mzIdentML. The main Perl parser reads results in CSV format, and uses additional config and parameters files to write valid mzIdentML (csv2mzIdentML for OMSSA; Tandem2CSV then csv2mzIdentML for Tandem). We have used the MSParser, produced by Matrix Science, for converting Mascot dat files to mzIdentML. File-parsing is carried out using separate Perl modules for each file format, so parsers for further file-formats may easily be added
- *Extract the PSMs from the mzIdentML files.* mzIdentML is converted to a tabular structure (or TSV file) for import into the re-scoring algorithms.
- *Perform the FDR analysis and, if required, combine FDR scores to give consensus FDR scores.* The FDR analysis algorithms¹¹ are encoded in Perl, reading from the TSV structures created in the previous stage. As such, by creating additional file adaptors or parsers, support for different search engines could be added in relatively simply.
- *Generate graphical output to assess the results.* Graphics are generated by a set of Perl library modules that take the results in the tabular structure and output png files.
- *Protein inference.* Proteins that share the same set (or a subset) of PSMs are assigned to the same group (encoded in mzIdentML within the ProteinAmbiguityGroup element). Individual protein identifications can also be ordered by a “Protein FDRScore” – calculated as the product of the individual FDRScores from each non-redundant PSM.
- *Report results for individual and combined search engines in both text and standard XML formats.* The results are exported from the re-scoring pipeline in CSV format and the csv2mzIdentML parser is used again to create a single mzIdentML file for the results after they have been combined from different search engines.

All of the components of the pipeline are freely available from a googlecode repository (<http://code.google.com/p/web-based-multiplesearch/>) so other developers can import all or part of them into their own software tools. Proteomics groups can also download a beta Java/Perl application to run on their own desktops that provides most of the pipeline functionality for those wishing to analyse large identification files or batches of files (the Mascot dat to mzIdentML feature is not currently available in the download due to licence restrictions).

Users interact with the website via the main search pages, which support selection and upload of results of searches from the Mascot, X!Tandem and/or OMSSA search engines.

Each of these files may be in either 'native' format ('.dat', '.xml' and '.csv' files, respectively) or in mzIdentML format. To benefit from the FDRAnalysis web tool, the search engines must have been run using a concatenated target/decoy database, resulting in one result file per search engine. Users must supply a 'tag' that the FDRAnalysis tool should use to identify decoy proteins. For example, if the decoy proteins are reverses of the 'correct' sequence, they might have the prefix 'REV_' to distinguish them. Users are also able to choose whether to wait for the analysis to complete or to receive an email notification upon completion. This is particularly useful when analysing large files (>10 MB), which may take some time to process. Control is also available as to whether to include N-terminal analysis, whether to combine the results from the different search engines, whether to produce graphical outputs, the size of the decoy database relative to the target database and the threshold FDR score for which results should be reported.

Results and Discussion

Summary statistics and download of results

When viewed through the web interface, FDRAnalysis provides the user with summary statistics in the form of a table displaying the number of target PSMs, the estimated number of true positives and the estimated number of false positives at 2 default FDRScore thresholds (0.01, 0.05) as well as a user selected threshold, if this is different to the defaults. The FDR calculation method is based on that described by Käll et al.⁷. If 'N-terminal analysis' was selected on the query form, the number of N-terminal PSMs are displayed in brackets. If a combined search using FDRScore was selected on the query form then these results will also be displayed in the table on the summary page. These results are produced by combining the results from the different search engines using the method described by Jones *et al*¹¹. The main summary page also contains hyperlinks, which allow users either to download result files or to view the graphs produced. An example summary page is shown in Figure 2.

The primary text file output available contains full details of the candidate peptide identifications, provided in a single TSV file containing a list of peptides identified by the combined FDR analysis. These are shown for different FDR score thresholds, preceded by similar summary statistics detailing the total number of identified peptides for each search engine at fixed FDRScore thresholds. An example (File 1) is shown in the Supplementary Data. Hyperlinks are also provided to download the results of the combined analysis in mzIdentML format. We believe this is an important addition to software tools in this area since it provides users with a means to convert native formats to PSI-sanctioned mzIdentML, as well as implementing the combined FDRscore approach. Currently, Mascot ('.dat' files), OMSSA ('.csv' files) and X!Tandem ('.xml' files) are supported. Due to the modular structure of the code, modules for parsing additional formats may easily be added in future. Additionally, via a BBSRC-funded project, an API for mzIdentML is under construction which will deliver format converters for most identification formats currently in use.

Graphical displays for assessing identifications

In addition to the raw data, we provide a set of informative plots in .png format containing graphs that illustrate the nature of the search results. The following graphs are produced by FDRAnalysis (Figure 3):

- Venn diagrams
- Rank distribution plots
- Score distribution plots

- Delta mass plots
- Start position plots

Two Venn diagrams can be displayed. An example of one of them is shown in Figure 3a. The first consists of peptides identified by the different search engines at the user-defined FDR. The second is displayed only if the user requests the combined FDRScore analysis. The combined approach nearly always leads to a larger number of PSMs at a given FDR threshold, by re-scoring identifications (in practice increasing the weighting given to identifications made by more than one search engine) which leads to a better separation between correct and incorrect identifications.

The rank plot generated by the FDRAnalysis tool enables the user to visualise the percentages of target and decoy hits at each rank (see figure 3b). For reliable FDR estimates, target and decoy peptides should be equally likely to be selected as matches under a random model. Hence, lower ranked PSMs (which are assumed to be incorrect) should have a ratio of target to decoy identifications equivalent to that used in the concatenated target/decoy database. The number of decoys hits has been adjusted by dividing by the decoy:target ratio of peptides within the database, as entered by the user. This should lead to an equal number of target and decoy hits for lower ranked PSMs. However, this may be affected by the choice of method for constructing decoy databases (such as the use of randomised sequences) and may lead to differing numbers of target and decoy peptides present in the *in silico* digested target and decoy proteomes^{2, 6}. On the other hand, the top ranking identifications above a certain threshold are expected to be correct and so should have a large bias towards target peptides. Consideration of both these phenomena can therefore be used to assess whether there are any inherent biases present in the decoy databases, as well as for individual search engines. Figure 3a shows a Mascot search result where the decoy database does not appear to have any such biases. The e-value cut-off for the top ranking peptides can be changed by the user on the query form (the default is 1.0), and the number of target and decoy hits at each rank are displayed for those spectra whose top ranking peptide surpasses the e-value threshold. Separate rank plots are shown for each of the search engines whose results are input by the user.

It is useful when considering score cut-off values for different search engines to have an idea of the number of target and decoy hits at different score thresholds. It would be expected that at lower scores, where peptides are assumed to be incorrect, there would be a target/decoy peptide hit ratio close to 1.0. At higher scores however, a large bias towards target peptides would be expected. Two different views show this distribution. The first is a simple plot of the percentage distribution of target and decoy peptides within different score ranges (figure 3c). The second is based on those described by Elias and Gygi⁶ and is shown in figure 3d. The number of incorrect identifications is estimated as the number of decoy peptides at a given score threshold divided by the ratio of decoy to target database size. The estimated number of correct identifications is calculated by subtracting the number of incorrect identifications from the number of target hits at a given score threshold. The OMSSA search engine does not provide an equivalent “raw” score, *per se*, and so the results are shown with respect to the expect values, transformed to $-\log(\text{expect})$ in order to give a distribution similar to that for the other search engines. As well as plots for the separate search engines, score plots are produced for the consensus peptides, if the ‘combined search’ option is selected. These plots display the distribution of peptides against their combined FDR values, rather than against a search engine specific score, again plotted on a $-\log$ scale.

Delta mass is defined as $(m/z_{\text{observed}} - m/z_{\text{theoretical}})$, where m/z is the mass-to-charge-ratio of a peptide’s product ions matched to the expected values for a given PSM. Two delta mass graphs give further valuable information concerning the performance of the different search

engines. For target identifications, we would expect PSMs with a high score to be clustered around a delta mass value of zero, indicating a small difference between theoretical and observed masses. On the other hand, decoy identifications are expected to have a much wider range of delta mass values¹⁶. These assumptions may be assessed using the first delta mass plot, which plots score against delta mass for the target and decoy PSMs (see figure 3e). The second delta mass graph shows whether there is a systematic bias in m/z assignments. It plots the percentage of PSMs that have a delta mass value within 0.1 of the given delta mass. This is expected to result in a normally distributed graph with a mean of zero (see figure 3f). However, in cases where there is an instrumental bias, the distribution may be centred on a non-zero value. This would indicate the need to make an adjustment to all m/z values in a pre-processing step before carrying out the searches and could be used to improve the calibration of the source instrument. Start position plots are useful for assessing the effectiveness of procedures for enriching in N-terminal and, to a lesser extent, C-terminal peptides. These plots are bar charts, which display one bar for the percentage of peptides that have been identified as coming from the N-terminus of their parent protein, i.e. have a start position of 1 or 2. The remaining positions are split into bins with the start position ranges 3-20, 21-50, 51-100 and 101+, with a bar displayed for each bin (see figure 3g).

Protein Inference

The software also performs basic protein inference from the PSMs, designed to be consistent with the PSI mzIdentML standard, using an algorithm that assigns proteins to groups if they share the same sets (or subsets) of PSMs. In mzIdentML there are structures to account for this ambiguity – called a ProteinAmbiguityGroup (PAG). A single (potential) protein identification in mzIdentML is called a ProteinDetectionHypothesis. These are grouped together within ProteinAmbiguityGroups. The FDRAnalysis tool calculates a basic score for each ProteinDetectionHypothesis, based on the product of the FDRScores calculated for each non-redundant PSM (i.e. taking the best score if the same peptide has been identified from multiple spectra) – similar to how protein scores are calculated in Mascot. The web server offers a download of output at the protein level, within a CSV file (ordered by product FDRScore) and encoded within a valid mzIdentML file.

Conclusions

FDRAnalysis is a software package that analyses the results from a number of different MS/MS search engines, based on decoy database searching. It accepts results in a number of different formats and outputs a list of candidate peptide and protein identifications, in mzIdentML, tab-separated and comma-separated formats. Internally, the program converts Mascot .dat, OMSSA .csv and tandem .xml files into mzIdentML compliant XML files. As a result, the program acts as a native-to-mzIdentML converter. In addition, the software can combine the results from different search engines to give a set of consensus results that have greater reliability and sensitivity than the results from any single search engine. FDRAnalysis includes a number of modules that allow visualisation of the data. The resulting graphs give users an improved understanding of the strengths and weaknesses of their results and of the relative performances of the different search engines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to gratefully acknowledge funding sources: ARJ, RK [BBSRC: BB/G010781/1] and SJH DCW, JAS [BBSRC: BB/F004605/1]. We are also grateful to David Creasy at Matrix Science for help with the MSParser.

References

1. Moore RE, Young MK, Lee TD. Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. *American Society for Mass Spectrometry*. 2002; 13:378–386.
2. Bianco L, Mead JA, Bessant C. Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. *Journal of Proteome Research*. 2009; 8(4):1782–1791. [PubMed: 19714810]
3. Bodenmiller B, Malmstrom J, Gerrits B, Campbell D, Lam H, Schmidt A, Rinner O, Mueller LN, Shannon PT, Pedrioli PG, Panse C, Lee HK, Schlapbach R, Aebersold R. PhosphoPep-a phosphoproteome resource for systems biology research in *Drosophila* Kc167 cells. *Molecular Systems Biology*. 2007; 3
4. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res*. 2008; 7(1):47–50. [PubMed: 18067251]
5. Dieguez-Acuna FJ, Gerber SA, Kodama S, Elias JE, Beausoleil SA, Faustman D, Gygi SP. Characterization of mouse spleen cells by subtractive proteomics. *Molecular & Cellular Proteomics*. 2005; 4(10):1459–1470. [PubMed: 16037072]
6. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007; 4(3):207–14. [PubMed: 17327847]
7. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7(1):29–34. [PubMed: 18067246]
8. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*. 2008; 7(1):40–4. [PubMed: 18052118]
9. Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal Chem*. 2004; 76(6):1664–71. [PubMed: 15018565]
10. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *Journal of Proteome Research*. 2008; 7(8):3354–3363. [PubMed: 18597511]
11. Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*. 2009; 9(5):1220–9. [PubMed: 19253293]
12. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20(18):3551–3567. [PubMed: 10612281]
13. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang XY, Shi WY, Bryant SH. Open mass spectrometry search algorithm. *Journal of Proteome Research*. 2004; 3(5):958–964. [PubMed: 15473683]
14. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem*. 2003; 75(4):768–74. [PubMed: 12622365]
15. McDonald L, Beynon RJ. Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nature Protocols*. 2006; 1:1790–1798.
16. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*. 2006; 24(10):1285–1292.

Synopsis

We present a web-based software tool (FDRAnalysis) that combines results from MS/MS spectra analysed using multiple search engines. FDRAnalysis calculates a combined score, based on false discovery rate, which results in more identifications and increased confidence in these identifications. The pipeline is based around the PSI-approved mzIdentML data standard and produces textual and graphical outputs that yield insights into the quality of the spectra and the search procedures.

Figure 1.
The FDRAnalysis pipeline. The boxed section contains the processing and analysis components. The remaining components comprise the web interface.

Figure 2.

The results summary page. This page is the main results page and is the first page displayed once the analysis is complete.

a.

b.

c.

Figure 3.

The graphs produced by FDRAnalysis: a. Venn diagram, showing the number of peptide hits passing a given FDR threshold made by single or multiple search engines; b. rank plot, showing the proportion of target and decoy PSMs at each rank; c. score distribution plot, showing the proportion of target and decoy PSMs at each search engine score; d. estimated numbers of correct/incorrect PSMs at different search engine scores, e. delta mass plot, displayed as search engine score vs. delta mass, f. delta mass distribution plot, showing the proportion of peptides with delta mass values within 0.1 of each delta mass value; g. start position plot, showing the distribution of the starting position of candidate peptide identifications with respect to the N-terminus of the parent protein.