003

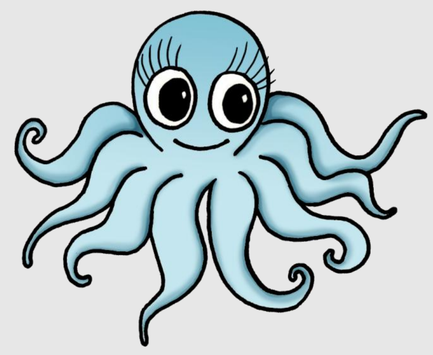
RUHR-UNIVERSITÄT BOCHUM

RUB

MEDIZINISCHES PROTEOM CENTER

Faculty of Medicine
Medizinisches Proteom-Center
WG Bioinformatics / Biostatistics

PIA - a toolbox for protein inference and identification analysis



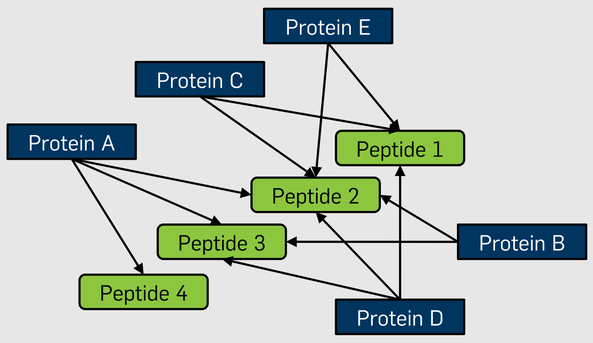
Julian Uszkoreit*, Helmut E. Meyer*, Christian Stephan*, Oliver Kohlbacher**, Martin Eisenacher*

*) Medizinisches Proteom-Center, Ruhr-Universität Bochum, **) Applied Bioinformatics Group, Eberhard Karls Universität Tübingen

Abstract

Most search engines for protein identification in MS/MS experiments return protein lists, although the actual search yields a set of peptide spectrum matches (PSMs). The step from PSMs to proteins is called “protein inference”. If a set of identified PSMs supports the detection of more than one protein in the searched database (“protein ambiguity”), usually only one representative accession is reported. These representatives may differ according to the used search engine and settings. Thus the protein lists of different search engines generally cannot be compared with one another. PSMs of complementary search engines are often combined to enhance the number of reported proteins or to verify the evidence of a peptide, which is improved by detection with distinct algorithms.
We introduce an algorithm suite written in Java, including a fully parametrisable web-interface (using JavaServer Faces on the server side), which combines PSMs from different experiments and/or search engines, and reports consistent and thus comparable results.
None of the parameters for the inference, like filtering or scoring, are fixed as in prior approaches, but held as flexible as possible, to allow for any adjustments needed by the user.

Challenges in Bottom-Up-Proteomics

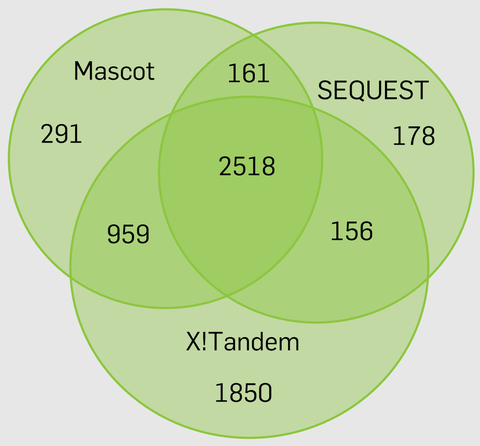


← Fig. 1: Peptide-protein ambiguity

After the theoretical digestion of the proteins most peptides are unique for one protein and suffice for the evidence of a protein (“Peptide 4” in the figure suffices for “Protein A”), but for some it is ambiguous to which protein of the used database a given peptide, and therefore a PSM, belongs (e.g. “Peptide 2” belongs to all the proteins in the figure).
To represent the connections between proteins and identified peptides and to speed up any further analysis, PIA creates a directed acyclic graph as internal data structure.

Fig. 2: Differing Spectrum Identifications →

Each peptide search engine uses its own algorithm to determine peptide spectrum matches (PSMs). Because of this, each search engine reports different PSMs for a given set of MS/MS spectra, consisting of either identifications also assigned by other algorithms or enhancing the list of possible assignments with new identifications. Sometimes an identification can even be contradictory to another identification. PIA gives the user the choice to use only PSMs for protein inference, which are e.g. identified by at least 2 of 3 search engines and above a given FDR threshold.
The adjacent Venn diagram shows the overlap of the PSMs found in a recent MS/MS dataset by the three search engines Mascot [1], SEQUEST [2] and X!Tandem [3].



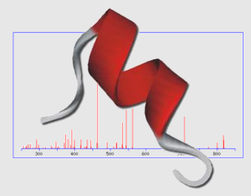
All	Occam‘s Razor	Needs unique
Protein A	Protein A	Protein A
Protein D	Protein D	
Protein B		
Protein C		
Protein E		

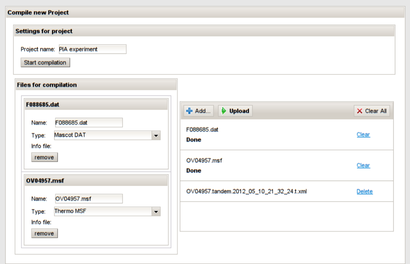
← Table 1: Protein report ambiguity

The adjacent table represents the results of one PSM identification run (the same as in figure 1) using different protein report strategies. While the strategy “All” simply reports all possible proteins, “Occam’s Razor” takes the smallest possible set of proteins explaining all the identified PSMs. At last, “Needs unique” only reports proteins having at least one peptide, which does not occur in any other protein of the database. Search engines use their own, sometimes adjustable, report strategies. PIA gives the user the opportunity to choose the report strategy and adjust its parameters.

Typical PIA workflow

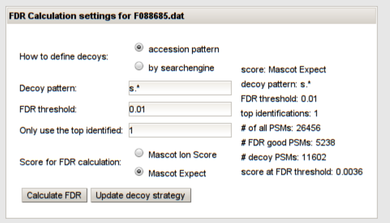
Step 1: First, the peptide database searches with any supported search engine are performed. Here the filter methods for e.g. “minimal peptide score” should not be too restrictive, as it can be further adjusted in later steps.





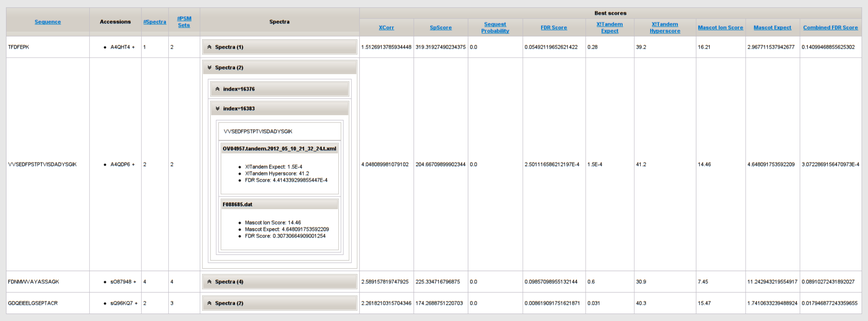
Step 2: In the “New Project” in the PIA web interface, which runs on any modern browser supporting JavaScript/AJAX, choose the search engine result files (runs) for the PIA experiment and start the compilation. At the moment, only local files are supported. A module for starting the searches with different search engines and automatically compiling the results is under development.
The compilation is time and memory consuming, but is needed only once per set of search engine results.

Step 3: The resulting compilation can now be browsed. The presentation is layered into three views (PSMs, peptides and proteins), which have own sets of basic analysis, filtering, sorting and exporting options.

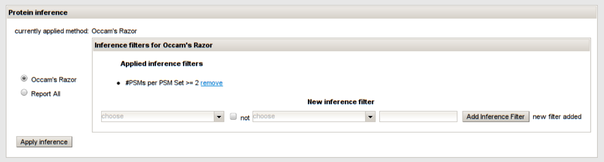


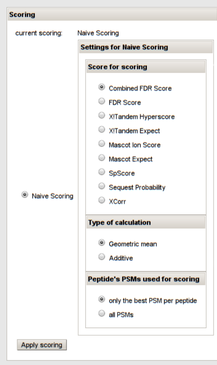
Step 3a: PSM Viewer
In this view, the PSMs for each run can be filtered and browsed. The false discovery rate (FDR) can be calculated and the PSMs are marked as valid or invalid, given an FDR threshold. Also, PSM sets are calculated, containing PSMs which come from the same spectra but from different runs. For a search engine independent metascore, the *FDR Score* [4] for each PSM set can be calculated.

Step 3b: Peptide Viewer
The peptides in this view are calculated either regarding the amino acid sequence only or also using the modification information of the PSMs. For each peptide, the sequence, accessions, spectra with further identification information and the best of each search engine score is shown. Filters like “need at least # of PSMs“ or “identified in specific run” can be set for the peptides.



Step 3c: Protein Viewer
After choosing an inference and scoring method, optionally with additional filters for the PSMs and peptides used, the resulting proteins can be browsed.





1) Perkins et al. (1999, electrophoresis), Probability-based protein identification by searching sequence databases using mass spectrometry data
2) Eng et al. (1994, Mass Spectrometry), An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database
3) Craig and Beavis (2004, Bioinformatics), TANDEM: matching proteins with mass spectra
4) Jones et al. (2009 Proteomics), Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines

Correspondence: Julian Uszkoreit, julian.uszkoreit@ruhr-uni-bochum.de

PURE

Julian Uszkoreit is funded by P.U.R.E. (Protein Unit for Research in Europe), a project of North Rhine-Westphalia, Germany