Journal of proteome research

Freie Universität Berlin

Subscriber access provided by UB + Fachbibliothek Chemie | (FU-Bibliothekssystem)

Technical Note

PIA - An intuitive protein inference engine with a web-based user interface

Julian Uszkoreit, Alexandra Maerkens, Yasset Perez-Riverol, Helmut E. Meyer, Katrin Marcus, Christian Stephan, Oliver Kohlbacher, and Martin Eisenacher

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.5b00121 • Publication Date (Web): 04 May 2015

PIA – An intuitive protein inference engine with a web-based user interface

Julian Uszkoreit, Alexandra Maerkens, Yasset Perez-Riverol ††, Helmut E. Meyer †+†††, Katrin Marcus, Christian Stephan †+††††, Oliver Kohlbacher †††††, Martin Eisenacher**

Medizinisches Proteom-Center, Ruhr-Universität Bochum, Bochum, Germany

KEYWORDS: protein inference, search engine combination, protein identification, peptide identification, identification analysis, database search, mass spectrometry

ABSTRACT. Protein inference connects the peptide-spectrum matches (PSMs) obtained from database search engines back to proteins, which are typically at the heart of most proteomics studies. Different search engines yield different PSMs and thus different protein lists. Analysis of results from one or multiple search engines is often hampered by different data exchange formats and lack of convenient and intuitive user interfaces. We present PIA, a flexible software suite for combining PSMs from different search engine runs and turning these into consistent results. PIA can be integrated into proteomics data analysis workflows in several ways. A user-friendly graphical user interface can either be run locally or (e.g., for larger core facilities) from a central server. For automated data processing, stand-alone tools are available. PIA implements several established protein inference algorithms and can combine results from different search engines

seamlessly. On several benchmark datasets we can show that PIA can identify a larger number of proteins at the same protein FDR when compared to inference based on a single search engine. PIA supports the majority of established search engines and data in the mzIdentML standard format. It is implemented in Java and freely available from https://github.com/mpc-bioinformatics/pia.

INTRODUCTION

In proteomics, the bottom-up or shotgun approach[1] has become the method of choice for high-throughput protein identification. The proteins of a sample are enzymatically digested to peptides and the resulting complex peptide mixture is then separated by liquid chromatography and typically analyzed using liquid chromatography coupled to mass spectrometry (LC-MS)[2]. The peptide ion mass spectra (MS) are acquired and fragment spectra (MS/MS) generated in a data-dependent fashion. The MS/MS spectra are either used for identification of peptides and proteins via database search engines like SEQUEST[3], Mascot[4], X!Tandem[5] or MS-GF+[6] or are identified by *de novo* peptide sequencing[7, 8].

As researchers are more often interested in the proteins rather than the peptides contained in an analyzed sample, most search engines and other workflows for MS/MS spectrum identification return protein lists containing database accessions, although the actual search determines peptide spectrum matches (PSMs). The step from the PSMs to proteins is called *protein inference*[9]. It is necessary, because a significant number of peptide hits are not unique but shared by different proteins in the database[10]. This is especially true for eukaryotic organisms due to homologous proteins or domains and multiple protein isoforms. These shared (or also called "degenerated" [9])

peptides lead to sets of proteins, called *protein ambiguity groups*, which are built up of the same (sub-)set of peptides and it cannot be decided which of the proteins were actually present in the sample unless discriminating (unique) peptides are found. Often for each such group only a representative accession number is reported in the result list and the other proteins are – if at all – reported as "similar proteins" or "group members". For a complete result list, all these possible proteins (according to the inference algorithm) should be reported as suggested by Nesvizhskii et al.[9] and is implemented in the mzIdentML[11] format.

The set of PSMs selected for protein inference, the inference algorithm, and the selection of reported representatives vary significantly between inference engines[12]. For some, mainly the algorithms included in commercial search engines and tools, but also some freely available algorithms, the details are scarcely described, so that results cannot be completely explained or it cannot be judged whether they are reasonable for a specific question. Additionally to search engine inherent inference algorithms, there are also stand-alone programs for protein inference from PSMs (e.g., ProteinProphet[13], Scaffold[14] and IDPicker[15]). Some of these support only specific search engines and most are limited in their settings for inference parameters[12].

Merging the results from multiple search engines is desirable to either increase the number of identified spectra passing an FDR threshold and thus hopefully also the number of corresponding proteins, or to solidify the evidence of peptides detected in the analyzed sample[16]. This poses a major problem, because each search engine's algorithm generates its own value for the quality of a PSM, generally a score or probability value (in the following score in this context always means the score or probability). These scores are usually not directly comparable. They thus need to be translated to a directly comparable, search-engine-independent score[17-19] prior to combining different search results.

In this work we present a set of algorithms and tools called "PIA – Protein Inference Algorithms" for the combination of PSMs obtained from different data sets and/or search engines. It reports consistent and comparable protein ambiguity groups as result of one of the implemented flexible protein inference methods. The implementation gives the choice of several protein inference and scoring methods and direct access to all required parameters. Essential analyses like the calculation of the false discovery rate (FDR) on the PSM level and the protein level are directly included. PIA is open-source software and completely written in Java. It provides an intuitive web–based graphical user interface (written in JavaServer Faces, JSF). This interface can either be used in local installation or via a public web server. The interface presents a fully filterable and browsable presentation and configuration of the steps from the PSMs and peptides to a protein list. For import and export, PIA supports the standard formats mzIdentML[11] and mzTab[20] for protein identifications developed by the HUPO Proteomics Standards Initiative (PSI[21]). For large-scale analyses are the use in central core facilities, PIA can also be called from the command line or embedded into the graphical workflow engine KNIME[22]. The latter also supports seamless integration into other MS/MS identification workflows, for example, workflows using OpenMS[23].

MATERIALS AND METHODS

Algorithms and implementation

The PIA algorithms and general workflow is based on three main concepts:

- A *peptide spectrum match* (PSM) refers to a match from an MS/MS spectrum to an amino acid sequence with charge state and identified modifications, which derives from one search engine run and contains the search engine's scores.

- A *peptide* in contrast refers to an amino acid sequence without charge state, either regarding modifications or not, depending on user settings used for the inference.
- A *protein* refers to an entry in a database (the raw amino acid sequence without any post-translational modifications), mandatorily containing an accession and, if available, a complete amino acid sequence and further descriptions.

The workflow in PIA is split into two main steps: data compilation of one or more search engine result files and the presentation and analysis of the data. The analysis occurs separately on all three identification levels (PSMs, peptides and proteins).

Compilation of result files: In this first step of PIA the PSM, peptide and protein data of all considered search engine runs is compiled into a directed acyclic graph which is stored into an intermediate XML file, together with additional search engine settings and identification information. This structured data (PIA intermediate structure) allows PIA or developers using PIA as a library fast access to the hierarchical information connecting all PSMs to peptides and proteins and vice versa (**Figure 1**) and provides an intuitive visualization of these connections. For each PSM the amino acid sequence, the associated protein(s), the charge state, the precursor *m/z* value, the difference between the measured and the theoretical peptide mass, the modifications and the retention time (if available) are collected. Additional protein information, like the complete protein sequence and human readable description, is also gathered and stored in the PIA XML's protein data.

After the data collection of all search engine runs the PSMs are assigned to their peptides, defined by their amino acid sequence. While doing so a map from the peptides to the proteins' accessions is built to accelerate subsequent lookups (**Figure 1 a**). Next, all PSMs and peptides in the map are structured into clusters, which form maximal connected sets with their mapped

proteins/accessions, i.e. all data in one cluster has no connection to any other cluster (**Figure 1 b**). These sets can subsequently be processed in parallel, to consecutively insert each peptide into its corresponding acyclic graph compartment along with its proteins. The graph is constructed in a straightforward way and consists of nodes for proteins, peptides with their PSMs and additional *group nodes* (**Figure 1 c**). The group nodes connect the protein and peptide nodes thus that the following rules are valid:

(1) each peptide and each protein belongs exactly to one group,

(2) a group can have other groups as children,

(3) there are no circles in the graph, even with respect to the (undirected) group-group relations,

(4) there is exactly one path from each protein to its peptides (with PSMs) and vice versa, which allows fast retrieval of the relations between proteins and peptides/PSMs.

After the compilation is finished, the graph data is stored in an XML file.

Analysis of identifications: This second step uses the compiled information of the first step. Results from multiple search engines for the same LC-MS run are assembled into *PSM sets,* which combine the identical identifications originating from different search engines. To assemble these sets, all basic PSM information (*m/z*, retention time, source ID, spectrum title, sequence, modifications and charge) available from all input files is used. If the assembly of PSM sets is not needed, it can be turned off, e.g. in case a compilation of successive LC-MS/MS runs is intended.

To evaluate the quality of the identification data and calculate the false discovery rate (FDR), a search against a target-decoy database is recommended[24, 25]. If such a search was conducted,

either a regular expression to distinguish decoy accessions from target accessions may be set, but also decoys generated by an internal target-decoy search can be used (e.g., used by Mascot and ProteomeDiscoverer). FDR, q-value and *FDR Score*[17] for each PSM are then calculated from this data. For PSM sets the *Combined FDR Score*[17] is computed as a comparable quality value for results from different search engines.

For an inspection of the data on the peptide level all PSMs and PSM sets with the same amino acid sequence are grouped into peptides. Additionally it can be specified, whether modifications should be considered to distinguish peptides. This peptide step can be used to review the peptides and associated PSMs of proteins of interest or to get a general overview of the identified peptides.

Protein inference algorithms: The protein inference in PIA depends on the choice of the method for the protein scoring, the inference algorithm and the setting of filters. Depending on the type of PSM scores (raw score, p-values or e-values) different rules for the protein scoring are applicable and can be chosen (e.g., addition, multiplication, geometric mean). For each rule one of the available PSM scores, including the *FDR Score* and the *Combined FDR Score*, may be chosen and it can be selected, whether all PSMs or only the best scoring PSM per peptide should be considered for the calculation of the protein score.

Due to shared peptides, homologs, isoforms, splice variants, or redundant database entries, it is often not possible to determine on the MS/MS data alone which proteins were truly present in the sample and thus should be reported. Therefore all current inference methods of PIA report protein ambiguity groups, which explain the same set of peptides, instead of single proteins. Though, depending on the settings and occurrence of unique peptides, a formal protein ambiguity group may contain a single accession only. According to the applied inference

method, a protein group may contain protein subgroups made up of subsets of the proteins' PSMs and/or peptides. For each inference algorithm, the PSMs and peptides, which should be used for the inference of the proteins, may be filtered to fulfill certain criteria, such as retaining an FDR level or contain at least two PSMs per peptide. These filters are the most important settings and facilitate the high configurability of PIA, besides the choice of the actual inference algorithm. Currently PIA implements three inference methods: (i) Report all, (ii) Occam's razor and (iii) Spectrum Extractor:

(i) *Report all*: This is the simplest possible inference method, just returning any possible protein group in the compilation of search results. Taking the PIA intermediate structure the reported proteins are very rapidly calculated, as only one protein group for each group in the graph containing at least one protein node needs to be created. The advantage of this method is its short runtime, with the disadvantage of calculating no sub proteins. This method does not report protein lists which would be accepted in current publications. But it can be used, to get a quick overview about the PSM and peptide data for a protein, which is actually not reported by any other method.

(ii) *Occam's razor*: Here the goal is to use the principle of maximum parsimony to report a minimal set of proteins, which explains the occurrence of all the identified peptides that pass the given filters. Given the example in **Figure 1** (and assuming no further filters) the protein groups with the single proteins A, D, H and I would get reported. This method also reports subgroups, in the example the group containing C and E would be a subgroup of D, the group with F and G a subgroup of H and the group J a sub group of I, whereas B's group would be a subgroup of both groups A and D.

(iii) *Spectrum Extractor*: The *Spectrum Extractor* is a spectrum-centric algorithm, in contrast to the two other implementations, which are peptide-centric. The major difference in this concept is that a spectrum, which gets once assigned to a peptide, never gets assigned to another peptide. This concept is closer to the reality, as in most cases one MS/MS spectrum contains only one peptide, although this may not always get the highest score by the search algorithms. This inference method is very similar although not equal to the inference method called "Protein Extractor"[26] implemented in the LIMS ProteinScape (Bruker, Bremen, Germany). If, instead of a score for a single PSM, a PSM set score (e.g., the *Combined FDR Score*) was selected as the base score for the inference, the combined PSM sets from multiple search engine runs are used for the inference.

The first step of this algorithm is the creation of a protein group for each group in the PIA intermediate structure containing any accession. Afterwards the following steps are performed:

1) For every protein group that has not yet reported examine each peptide. If a peptide is already reported, allow it to be reported in this protein group with the prior set PSMs and score. Else, construct the peptide with all still available PSMs fulfilling the inference filters.

If a spectrum is present in more than one peptide in a protein group, use it for the protein scoring only in the peptide, where it has the best score.

Should there be more than one peptide in a protein group, where the spectrum has the best score, collect all spectra, which may count for the affected peptides. If there are peptides, which have all of the affected spectra, one of these peptides is used with all the spectra while scoring, all other peptides are not considered during scoring. If the affected spectra are distributed over several peptides, calculate the score of these peptides without the questionable spectra. The peptide with the best score gets all its spectra assigned. If there are peptides with the same score

and spectra, all of them get the spectra assigned but only one is considered for the protein scoring. Repeat these last steps, until all spectra are assigned to peptides.

2) Calculate the score for each protein group and select the group with the best score. Check, whether this protein group is a subgroup of any already reported protein group regarding peptides or PSMs. If it is a same-set (i.e. the protein groups contain the same PSMs and peptides) or sub protein group, assign it to the respective group appropriately. If not, add the protein group and all its peptides and PSMs to the set of reported items and report this protein group.

3) Repeat steps 1 and 2 until there are no further protein groups to be reported

Implementation and data representation. PIA is developed in Java and all its components can be used directly from the command line; thus it can be integrated into any scripted identification pipeline. A more user-friendly way of using PIA is the web interface. The presentation in the web interface as well as the analysis steps are layered into three levels, corresponding to PSMs, peptides and proteins. The default and most intuitive procedure of an analysis using the web interface is the wizard (**Figure 2**). After the compilation is finished, the wizard assists a user through the default steps of an analysis by performing an FDR calculation on PSM level, choosing a protein inference and scoring method and performing the inference. Additionally, after each step some descriptive statistics are shown, on which, for example, a basic quality check can be performed. The wizard can be aborted at any time, which directs the user to a more advanced interface (**Figure 3**), where all steps and calculations can be performed by direct user request and any interim results can be reviewed immediately. The advanced interface allows also for an in-depth inspection of the identification results, the results of the combination from different search runs and the inferred peptides from the (combined) PSMs. For filtering and exporting the PSM, peptide and protein lists, the user can select a variety of

variables and thus filter for score, mass deviation, sequence, or other attributes. Furthermore, an intuitive visualization of the relations of the PSMs, peptides and accessions, which lead to the reported protein groups, was implemented. For this purpose the part of the PIA intermediate structure leading to a given protein is depicted. In the resulting image, the peptides leading to the protein and the occurring sub-proteins are highlighted, as shown in **Figure 4**. The web interface is written for JavaServer Faces (JSF) which needs a running installation of a JavaServer Pages web server (e.g., Apache Tomcat[27] or GlassFish Server[28]). The interface may then be accessed via any current browser either locally, via a network or via the Internet from any modern computer. A deployable binary version including all needed dependencies for the web server is available for download on the project homepage, as well as a link to the public demo server.

To integrate PIA into new or existing KNIME workflows, nodes are developed using the GKN[29, 30] package and can be downloaded on the project homepage. This allows an integration of PIA into larger pipelines (e.g., using OpenMS).

Supported data formats. The current implementation supports importing mzIdentML[31] and thus import from virtually all search engines, provided a converter into this standard format exists already or the search engine directly exports into mzIdentML, like MS-GF+. Alternatively the import of idXML files generated by OpenMS, but also a convenient import of search engine native Mascot DAT files and X!Tandem XML is supported, using the open source parsing tools described elsewhere[32, 33]. Additionally an importer for data from ProteomeDiscoverer 1.3 and 1.4 files is implemented and tested for Mascot, SEQUEST (default and HT version) and MS Amanda[34] searches.

The standard format of the PSI for reporting peptide and protein identifications, mzIdentML, facilitates the comprehensive reporting of protein ambiguity groups and their members[35] and is

therefore the format of choice for exporting protein data from PIA. An exporter into the less complete but easier human readable and computer parsable PSI format mzTab or a simple comma separated values (CSV) format is also implemented.

Datasets

To evaluate the reliability and to describe the behavior of PIA, it was assessed on one real-life in-house dataset, of which the precise protein contents are not known, and on two public datasets with knowledge of the protein contents. The in-house dataset is a label-free mass spectrometry analysis of a murine cell culture sample. The first dataset with known protein content is part of the "Gold Standard of Protein Expression in Yeast" also used by Ramakrishnan et al.[36]. The other dataset also containing known proteins was produced for the Proteome Informatics Research Group (iPRG) 2008 study of the Association of Biomolecular Resource Facilities (ABRF). These datasets were used to measure and compare the performances of the PIA algorithms using the common search engines Mascot (Version 2.4.1), MS-GF+ (v9949) and X!Tandem (Sledgehammer - 2013.09.01.1). PIA intermediate XML files were generated with various search engine's result files per dataset and used to generate protein group results with different PIA settings and filters. The generation of the intermediate files and the report lists were performed on a laptop computer and took less than 1.5 hours per dataset.

All benchmarking datasets, the plotted search results and used KNIME workflows have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[37] with the dataset identifiers PXD000790, PXD000792, PXD000793.

Mouse Dataset. For the creation of the mouse dataset cultured cells of a murine myoblast cell line were harvested and centrifuged for 5 min at 800 x g. The cell pellet was resuspended in lysis buffer (3 mM Tris-HCl, 7 M urea, 2 M thiourea, pH 8.5), homogenized and lysed via sonification (6 times for 10 sec, on ice). After centrifugation (15 min, 16,000 x g) the supernatant was collected and protein content was determined by Bradford protein assay. For the following tryptic in-solution digestion 20 µg of sample was diluted in 50 mM ammoniumbicarbonate (pH 7.8) to a final volume of 100 µl, reduced by adding DTT and alkylated with iodacetamide as described previously in [38]. After digestion the peptide concentration was determined by aminoacid analysis and 200 ng of the peptide sample was subsequently analyzed by a label-free mass spectrometry approach using an UltiMate 3000 RSLC nano LC system directly coupled to an LTQ Orbitrap Elite mass spectrometer (both Thermo Fisher Scientific, Dreieich, Germany, the protocol is further described in supplemental file 4).

For spectrum identification an mzML file was created from a Thermo RAW file using the msConvertGUI of ProteoWizard[39] and further converted into an MGF file by OpenMS. This MGF was searched against a decoy database of the *Mouse Complete Proteome Set* downloaded from UniProtKB on 26.11.2014 (44,467 entries). A shuffled decoy database was created with the DecoyDatabaseBuilder[40]. The search engines used a parent mass tolerance of 5 ppm, fragment mass tolerance of 0.4 Da and allowed one missed cleavage. Oxidation of M, acetylation of the protein N-terminus, Glu to pyro-Glu and Gln to pyro-Glu were used as variable modifications, carbamidomethylation of C as fixed modification.

Yeast Gold Standard Dataset. The RAW data files were downloaded from http://www.marcottelab.org/MSdata/Data_02. The measured samples contain proteins of wild-type yeast, growing in rich medium, harvested in the log phase. The expressed proteins contained

in the sample were identified by MS- and not-MS-based methods and are available as a reference set. For the performance measurement on this dataset, a shuffled decoy database of the current version of the protein database from the *Saccharomyces Genome Database* (SGD, www.yeastgenome.org, downloaded on 28.05.2014, 6,717 entries) was used for protein identification. As some of the entries in the reference set are no longer in the SGD database due to newer protein annotations, the reference set of proteins to be known in the sample was adjusted (for more information see supplemental file 1) and finally contains 4,258 accessions. Of the original 32 RAW files (eight different mass spectrometer settings with four SCX salt steps each) available, the four runs of the mass spectrometer settings with the most spectra were used (070119-zl-mudpit07-1). For these runs the RAW files were converted to mzML using the msConvertGUI and further processed to MGF files using OpenMS. For the identification a precursor tolerance of 25 ppm, a fragment tolerance of 0.5 Da, one missed cleavage and the variable modifications for oxidation of M and protein N-terminal acetylation were allowed.

iPRG2008 dataset. The used MGF files and the concatenated target-decoy database were downloaded from the homepage of the iPRG (http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm#786). These data was also provided for the ABRF iPRG208 Study which aimed at the goal to "assess the quality and consistency of protein reporting on a common data set", as stated on the studies slides. For this study, mouse samples were trypsin digested and the peptides labeled by four-plex iTRAQ and fractionated via strong cation exchange chromatography. The fractions were measured by LC-MS/MS on a 3200 QTrap, some fractions multiple times with different exclusion lists, which resulted into 29 files. These data was analyzed by members of the iPRG by a variety of search engines and inference tools. The results were used to create a list of protein

clusters, which are detectable in the data. One protein cluster contains multiple accessions, which share some peptide information. For each cluster the number of expected identifications was identified using the iPRG's members analyses. Furthermore, the clusters were assigned to 5 different classes, but only the first three classes were graded in the further assessment. Class 1 (16 clusters) contains consensus multiple identifications, class 2 (11 clusters) contains debatable multiple identifications and class 3 (182 clusters) contains consensus single identification per cluster. For more information please consult the iPRG's homepage.

For the peptide identification a precursor and fragment tolerance of 0.45 Da and one missed cleavage were allowed. For the fixed modifications four-plex iTRAQ on K and N-termini as well as methylthio on C and for the variable modifications oxidation of M and protein N-terminal acetylation was used.

RESULTS AND DISCUSSION

Mouse Dataset. On this dataset the application of PIA on a current dataset was assessed. For this, the searches performed by Mascot, MS-GF+ and X!Tandem were first analyzed separately and then a combination of all searches. The numbers of identified protein groups using the "Spectrum Extractor" are plotted against the protein FDR q-value in **Figure 5**. For Mascot searches three protein inferences were performed using allowed PSM FDR Score values below 0.30, below 0.10 and below 0.01 (the recommended value by the authors) , respectively. While decreasing the allowed FDR level also decreases the total number of reported proteins, the number of target proteins in the low FDR range is increased, i.e. the beginning of the list contains fewer false positives. This increase of reported high quality proteins is only observable until a certain FDR level is reached, below which the number of reported proteins rapidly

decreases, as can be seen in the plot when allowing only PSMs up to an FDR below 0.01. Additionally the numbers of proteins when only using identifications from MS-GF+ respectively X!Tandem and FDR below 0.01 are plotted, which show equal trends though different numbers of reported proteins at given q-values. Finally, the number of reported proteins when using the combination of all search engines and keeping the PSM FDR level (using the *Combined FDR Score*) at 0.01 exceeds the number of reported proteins for each single search engine at every q-value. This indicates that a combination of search engine results with PIA improves the number of true identifications in a list of protein groups.

Yeast Gold Standard Dataset

The performance of PIA using the Spectrum Extractor and Occam's Razor inference, with the need for one unique peptide per reported protein group, was analyzed for each search engine and the combination of search engines on this dataset. As for this dataset the proteins contained in the sample are known, the local FDR and q-value of the ranked protein results can be calculated using the proteins contained in the reference set as true positive identifications and all other identifications as false positives. With these values a pseudo ROC curve plotting the number of true positives against the corresponding q-values depicts the quality of the results. In **Figure 6** the curves for the combination by PIA and the X!Tandem results alone with at least one unique peptide per protein group are shown (lists for all searches are contained in the supplemental file 2). Though the general behavior is similar, it is interesting to note, that for the used dataset the Spectrum Extractor usually yields better performance in the very low q-value regions but the overall number of reported proteins is usually higher with Occam's Razor. Though these observations are dataset dependent, the data show overall good results for the inference algorithms used and does not make many false reports, as the plotted curves all stop before the

value of 0.035. For all analyzed settings, the protein group with the accessions YLR227W-B and YPR158C-D was identified at around rank 60, although it is not in the reference set. The quality of the identification though indicates that it is a false negative. Usually it can be said, that the Spectrum Extractor reports fewer proteins because it uses a spectrum only for one peptide, if the search engine reports more than one PSM per spectrum. Again, the combination of search results by PIA yields more highly evident protein groups.

iPRG2008 dataset. In this dataset the expected number of identifications per cluster was calculated by the ABRF group members. For the classes 1, 2 and 3 these numbers are assessed and result in a total maximum of 258 true positive (TP) identifications. A false positive (FP) identification has too many identifications per cluster, whereas a false negative (FN) identification has too few identifications. In **Figure 7** the results of the (a) totally reported, (b) TP, (c) FN and (d) FP identifications are shown for protein inferences conducted by PIA in comparison to the average outcome of the iPRG2008 study. With PIA the PSMs with FDR below 0.01of each search engine alone and in combination were inferred to a protein group list using the Spectrum Extractor, for the comparison only proteins with protein FDR below 0.01 were used. The combination of search results yields the highest number of reported proteins and also outperforms most of the iPRG study participants, only 4 of 23 participants reported more. More interestingly, also the number of true positives is much higher in the report of the combination, which is surpassed by only 6 iPRG participants (compare lists and charts in the sheet “overview2” of supplemental file 3). The false negatives rates with the assessed PIA settings are better than the average iPRG participant’s results. An exception is formed by the MS/GF+ search, which reports the fewest proteins at all and therefore also the highest false negatives, while the PIA combination is outperformed by only 6 of the iPRG participants. The

relatively high numbers of false positives in all runs except the MS-GF+ run correspond mainly to clusters, which are also dubious in the slides of the ABRF study (compare also supplemental file 3). For these, many ABRF group members and study participants found more than the expected number of distinguishable detectable isoforms. The number of false positives can be decreased by stricter inference parameters like the need to have at least one unique peptide per protein. Though, stricter settings also decrease the number of totally reported proteins and thus true positives.

For the datasets with known contents, plots of the target-decoy estimated protein level q-values against the (claimed) true protein level q-values are shown in supplemental file 5. These plots show significant differences between the datasets. For the iPRG2008 dataset, the estimated error is consistently much lower than the actual value, though the ratio goes down with the number of reported proteins. For the Yeast Gold Dataset, the actual values are underestimated on the top of the protein list and overestimated after a certain value (for the combination of all search engines at an FDR of 0.03). These differences are presumably due to the underlying ways, of how the actual protein content was measured. For the iPRG2008 dataset prior search results of the same actual MS data were used and thus identifying more proteins, than the proteins which are claimed to be valid is more probable with different search engines. For the determination of the yeast dataset's content, also other technologies were used, which allows to create a more complete compilation of the contained proteins. For an in-depth analysis of the estimation between the true and estimated protein q-values of protein inference algorithms, more datasets of complex protein mixtures with exactly known content would be needed, which are not available at the time of writing this manuscript.

CONCLUSIONS

In this paper we introduced PIA, a new toolbox for protein inference and identification analysis, to solve some common protein ambiguity problems of LC-MS/MS proteomics and improve the results of identification experiments. Different protein inference and scoring methods can be quickly tested and their results be compared, either by using a browser based user friendly interface which also allows for an in-depth analysis, or using command line tools or KNIME nodes for pipeline environments.

Although other tools for the combination of search results, estimation of identification quality and inference of proteins from peptide identifications exist, most of them give only very few adjustable settings. With PIA we allow the user to adjust almost all settings to the desired needs. The reported PSM, peptide and protein lists can easily be inspected to find the responsible peptides and even PSMs for any reported protein. An export into easily parseable generic formats (CSV and mzTab) or more advanced formats like mzIdentML is included for further processing. With the import of mzIdentML we support virtually any search engine without the need of further implementations, while also native result files of some of the most used search engines can be imported – at the time of writing Mascot, X!Tandem and all ProteomeDiscoverer 1.3 and 1.4 results.

The analyzed data show that PIA returns valid protein groups for well characterized datasets and also emphasizes the usage of more than one search engine for the analysis of mass spectrometry data to improve the results. Parts of PIA, especially to perform fast and comparable spectral counting analyses, were already used in recent publications ([38, 41, 42]) and prove the applicability of the results. Algorithms of PIA are included in the recent version of PRIDE Inspector[43] to allow protein inference and visualization using the ms-data-core-api[44]. Future tasks

for the development are the implementation of further protein inference and scoring algorithms to better discriminate isoforms, make a decision of a protein group's representative accession, and incorporation of quantitation data.

ASSOCIATED CONTENT

Supporting Information. Detailed information about the generation of the SGD database (supplemental file 1), analysis of the yeast gold dataset (supplemental file 2) and iPRG2008 dataset (supplemental file 3), detailed description of the generation of the mouse dataset (supplemental file 4), comparison of estimated vs. calculated q-values (supplemental file 5). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: martin.eisenacher@ruhr-uni-bochum.de. Phone: +49-234-32-29288. Fax: +49-234-32-14554

*E-mail: julian.uszkoreit@ruhr-uni-bochum.de. Phone: +49-234-32-29275

Present Addresses

† Medizinisches Proteom-Center, Ruhr-University Bochum, Bochum, Germany

†† European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

††† Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund, Germany

†††† Kairos GmbH, Bochum, Germany

††††† Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center, and Dept. of Computer Science, University of Tübingen, Tübingen, Germany

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

J.U and M.E are funded by PURE (Protein Unit for Research in Europe), a project of North Rhine-Westphalia, Germany. A.M is funded by the German Research Foundation (FOR1228). Y.P.R is supported by the BBSRC 'PROCESS' grant [reference BB/K01997X/1].

Notes

PIA is released under a three-clause BSD license and freely available for download at https://github.com/mpc-bioinformatics/pia.

ACKNOWLEDGMENT

The authors thank Britta Eggers for processing and measuring of the murine cell culture samples on the mass spectrometers. We also thank the federal state North Rhine-Westphalia for funding within the project PURE (Az.: 131/1.08-031). The data deposition to the ProteomeXchange Consortium was supported by PRIDE Team, EBI.

ABBREVIATIONS

PSM, Peptide Spectrum Match; FDR, False Discovery Rate; CV, controlled vocabulary; PSI, HUPO Proteomics Standards Initiative; iPRG, Proteome Informatics Research Group; ABRF, Association of Biomolecular Resource Facilities

REFERENCES

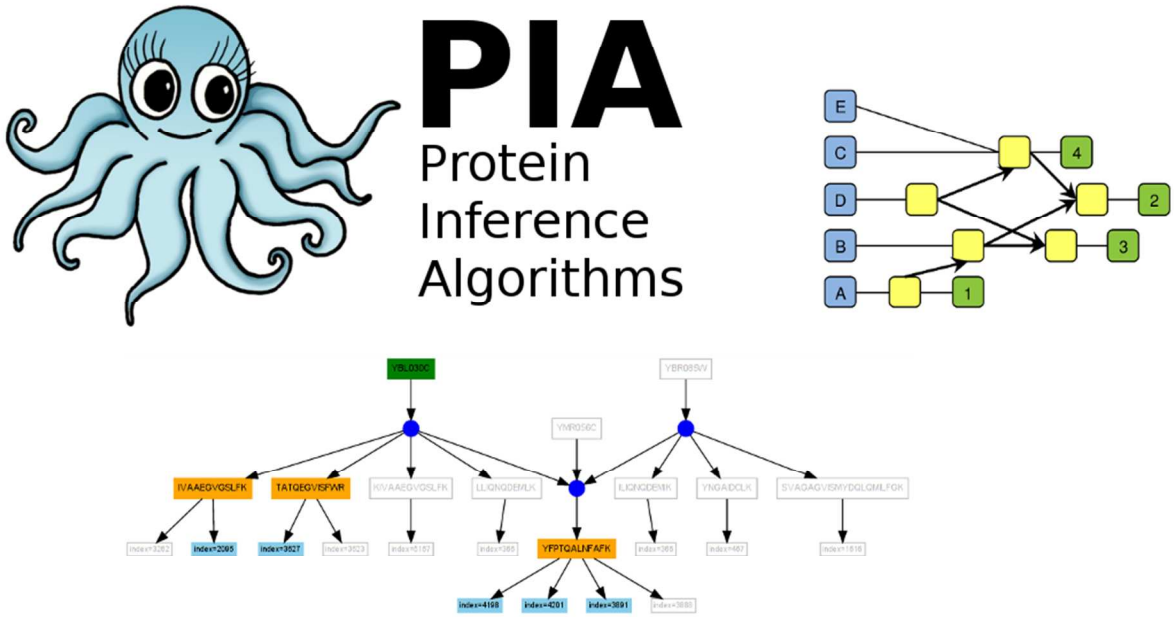
1. Wolters, D. A.; Washburn, M. P.; Yates, J. R., An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* **2001,** 73, (23), 5683-90.
2. Perez-Riverol, Y.; Wang, R.; Hermjakob, H.; Müller, M.; Vesada, V.; Vizcaíno, J. A., Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim Biophys Acta* **2014,** 1844, (1 Pt A), 63-76.
3. Eng, J.; McCormack, A.; Yates, J., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.
4. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-67.
5. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.
6. Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J.; Pevzner, P. A., The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol Cell Proteomics* **2010,** 9, (12), 2840-52.
7. Bandeira, N., Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques* **2007,** 42, (6), 687, 689, 691 passim.
8. Bertsch, A.; Leinenbach, A.; Pervukhin, A.; Lubeck, M.; Hartmer, R.; Baessmann, C.; Elnakady, Y. A.; Müller, R.; Böcker, S.; Huber, C. G.; Kohlbacher, O., De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis* **2009,** 30, (21), 3736-47.
9. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005,** 4, (10), 1419-40.
10. Perez-Riverol, Y.; Sánchez, A.; Ramos, Y.; Schmidt, A.; Müller, M.; Betancourt, L.; González, L. J.; Vera, R.; Padron, G.; Besada, V., In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J Proteomics* **2011,** 74, (10), 2071-82.
11. Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P. A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D., The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* **2012,** 11, (7), M111.014381.

12. Huang, T.; Wang, J.; Yu, W.; He, Z., Protein inference: a review. *Brief Bioinform* **2012,** 13, (5), 586-614.
13. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** 75, (17), 4646-58.
14. Searle, B. C., Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010,** 10, (6), 1265-9.
15. Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L., IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **2009,** 8, (8), 3872-81.
16. Eisenacher, M.; Kohl, M.; Turewicz, M.; Koch, M. H.; Uszkoreit, J.; Stephan, C., Search and decoy: the automatic identification of mass spectra. *Methods Mol Biol* **2012,** 893, 445-88.
17. Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W., Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **2009,** 9, (5), 1220-9.
18. Nahnsen, S.; Bertsch, A.; Rahnenführer, J.; Nordheim, A.; Kohlbacher, O., Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *J Proteome Res* **2011,** 10, (8), 3332-43.
19. Kwon, T.; Choi, H.; Vogel, C.; Nesvizhskii, A. I.; Marcotte, E. M., MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. *J Proteome Res* **2011,** 10, (7), 2949-58.
20. Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q. W.; Del Toro, N.; Perez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaino, J. A.; Hermjakob, H., The mzTab Data Exchange Format: communicating MS-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* **2014**.
21. Orchard, S.; Hermjakob, H.; Apweiler, R., The proteomics standards initiative. *Proteomics* **2003,** 3, (7), 1374-6.
22. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications* **2008**, 319-326.
23. Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O., OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008,** 9, 163.
24. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
25. Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **2008,** 7, (1), 29-34.
26. Korting, G.; Bluggell, M.; Marcus, K.; Chamrad, D. C.; Lohaus, C.; Reidegeld, K.; Stephan, C.; Schweiger-Hufnagel, U.; Glandorf, J.; Meyer, H. E.; Thiele, H., Protein extractor; from peptide ID to protein ID. *Molecular & Cellular Proteomics* **2006,** 5, (10), S216-S216.
27. Apache Tomcat. http://tomcat.apache.org
28. GlassFish Server. https://glassfish.java.net/

29. de la Garza, L.; Krüger, J.; Schärfe, C.; Röttig, M.; Aiche, S.; Reinert, K.; Kohlbacher, O., From the Desktop to the Grid: conversion of KNIME Workflows to gUSE. In *Proceedings of the 5th International Workshop on Science Gateways*, Kiss, T., Ed. 2013.
30. GenericKnimeNodes. https://github.com/genericworkflownodes/GenericKnimeNodes (19.09.2014),
31. Reisinger, F.; Krishna, R.; Ghali, F.; Ríos, D.; Hermjakob, H.; Vizcaíno, J. A.; Jones, A. R., jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. *Proteomics* **2012,** 12, (6), 790-4.
32. Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K., MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* **2007,** 7, (3), 364-6.
33. Muth, T.; Vaudel, M.; Barsnes, H.; Martens, L.; Sickmann, A., XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* **2010,** 10, (7), 1522-4.
34. Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K., MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res* **2014,** 13, (8), 3679-84.
35. Seymour, S. L.; Farrah, T.; Binz, P. A.; Chalkley, R. J.; Cottrell, J. S.; Searle, B. C.; Tabb, D. L.; Vizcaíno, J. A.; Prieto, G.; Uszkoreit, J.; Eisenacher, M.; Martínez-Bartolomé, S.; Ghali, F.; Jones, A. R., A standardized framing for reporting protein identifications in mzIdentML 1.2. *Proteomics* **2014**.
36. Ramakrishnan, S. R.; Vogel, C.; Prince, J. T.; Li, Z.; Penalva, L. O.; Myers, M.; Marcotte, E. M.; Miranker, D. P.; Wang, R., Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics* **2009,** 25, (11), 1397-403.
37. Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Pérez-Riverol, Y.; Reisinger, F.; Ríos, D.; Wang, R.; Hermjakob, H., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* **2013,** 41, (Database issue), D1063-9.
38. Kley, R. A.; Maerkens, A.; Leber, Y.; Theis, V.; Schreiner, A.; van der Ven, P. F.; Uszkoreit, J.; Stephan, C.; Eulitz, S.; Euler, N.; Kirschner, J.; Müller, K.; Meyer, H. E.; Tegenthoff, M.; Fürst, D. O.; Vorgerd, M.; Müller, T.; Marcus, K., A combined laser microdissection and mass spectrometry approach reveals new disease relevant proteins accumulating in aggregates of filaminopathy patients. *Mol Cell Proteomics* **2013,** 12, (1), 215-27.
39. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** 30, (10), 918-20.
40. Reidegeld, K. A.; Eisenacher, M.; Kohl, M.; Chamrad, D.; Körting, G.; Blüggel, M.; Meyer, H. E.; Stephan, C., An easy-to-use Decoy Database Builder software tool, implementing

different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. *Proteomics* **2008,** 8, (6), 1129-37.
41. Schrötter, A.; Mastalski, T.; Nensa, F. M.; Neumann, M.; Loosse, C.; Pfeiffer, K.; Magraoui, F. E.; Platta, H. W.; Erdmann, R.; Theiss, C.; Uszkoreit, J.; Eisenacher, M.; Meyer, H. E.; Marcus, K.; Müller, T., FE65 regulates and interacts with the Bloom syndrome protein in dynamic nuclear spheres - potential relevance to Alzheimer's disease. *J Cell Sci* **2013,** 126, (Pt 11), 2480-92.
42. Maerkens, A.; Kley, R. A.; Olivé, M.; Theis, V.; van der Ven, P. F.; Reimann, J.; Milting, H.; Schreiner, A.; Uszkoreit, J.; Eisenacher, M.; Barkovits, K.; Güttsches, A. K.; Tonillo, J.; Kuhlmann, K.; Meyer, H. E.; Schröder, R.; Tegenthoff, M.; Fürst, D. O.; Müller, T.; Goldfarb, L. G.; Vorgerd, M.; Marcus, K., Differential proteomic analysis of abnormal intramyoplasmic aggregates in desminopathy. *J Proteomics* **2013**.
43. Wang, R.; Fabregat, A.; Ríos, D.; Ovelleiro, D.; Foster, J. M.; Côté, R. G.; Griss, J.; Csordas, A.; Perez-Riverol, Y.; Reisinger, F.; Hermjakob, H.; Martens, L.; Vizcaíno, J. A., PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* **2012,** 30, (2), 135-7.
44. Perez-Riverol, Y.; Uszkoreit, J.; Sanchez, A.; Ternent, T.; del Toro, N.; Hermjakob, H.; Vizcaíno, J. A.; Wang, R., ms-data-core-api: An open-source, metadata-oriented library for computational proteomics. *Bioinformatics* **2015**.

Table of Contents Graphic



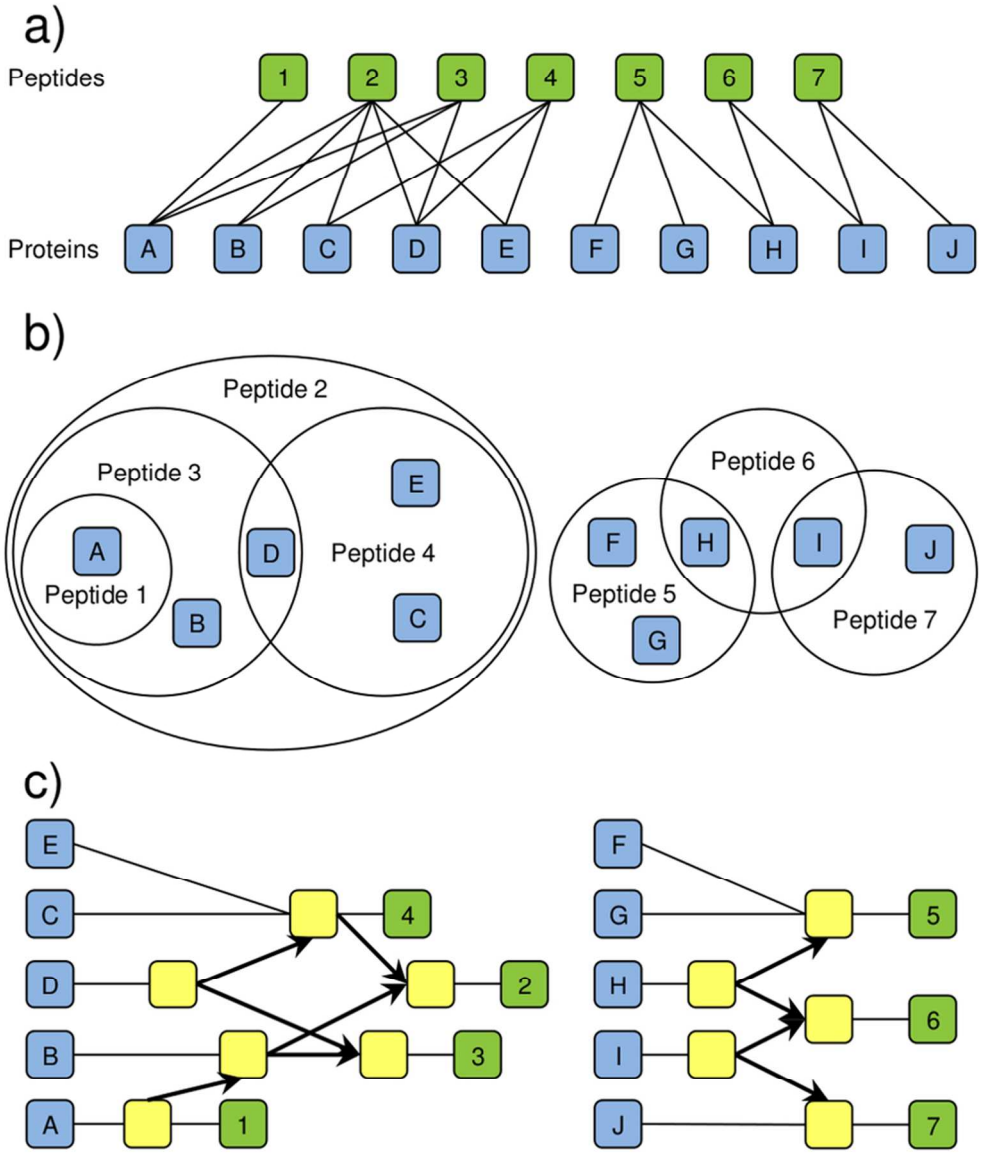


Figure 1. The compilation of the search engine results into a directed graph is performed in three steps. The PSMs can easily be grouped into peptides according to their amino acid sequences, therefore PSMs are left out in this figure. The connection information between peptides (green) and proteins (blue) is stored in a map shown in a), where each peptide belongs to one or more proteins. This map can be divided into closed clusters, where each peptide maps to all its proteins and there is no mapping from one cluster to any other, as depicted by two such clusters in b). The information of these closed clusters can be processed in parallel to create a set of acyclic graphs shown in c), where it is easy to retrieve for a given protein all peptides (and PSMs) or vice versa by following the connections between nodes. This data structure is the actual used intermediate format used by PIA to quickly retrieve information. The yellow "group" nodes store no additional information, but are necessary to connect the remaining nodes correctly and uphold the set of rules given in the text (compilation of result files).
83x101mm (300 x 300 DPI)

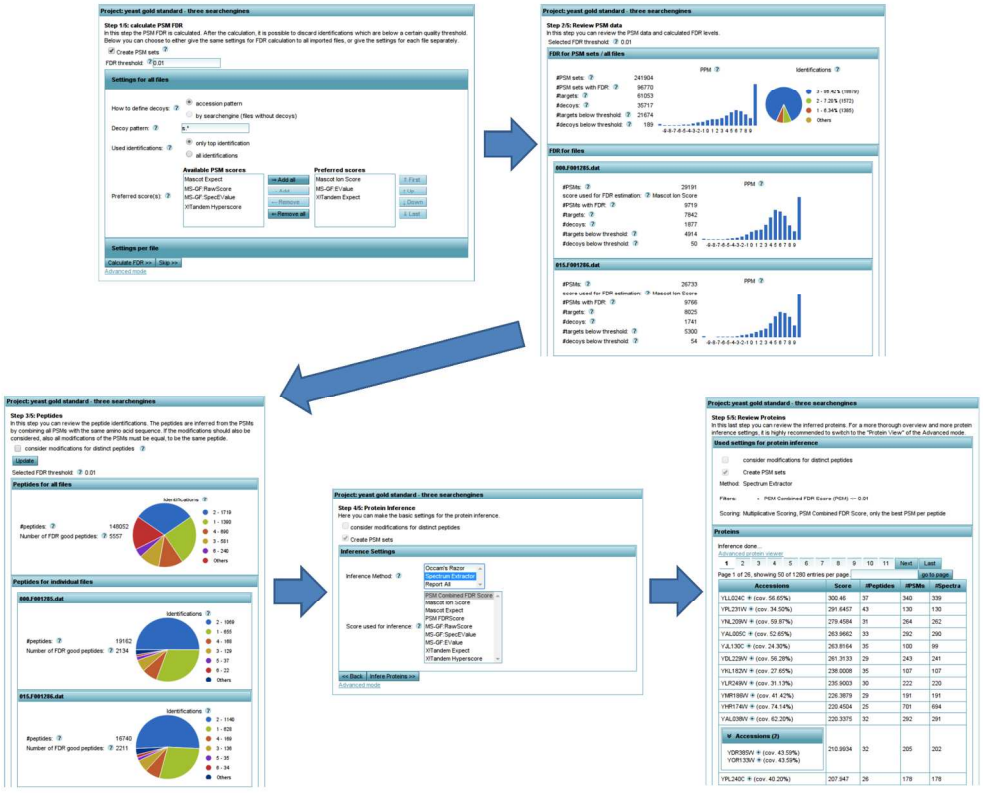


Figure 2. The wizard of the PIA web interface. These screenshots demonstrate the steps of the PIA wizard, the most convenient way to perform an analysis. The wizard is part of the web interface and guides the user through the analysis while suggesting defaults for most of the used parameters. The wizard starts with the calculation of the PSM FDR values for each search engine run (step 1) and shows statistics on these values like the number of target and decoy PSMs as well as the distribution of mass deviations (step 2). In the third step, the PSMs are inferred to peptides and an overview of the number of identifications per peptide is shown. In step four the protein inference method is selected and processed. The final step 5 shows a short overview of the inferred proteins. The visualization in the wizard can be aborted at any time and instead be switched to the advanced viewer.
176x144mm (300 x 300 DPI)

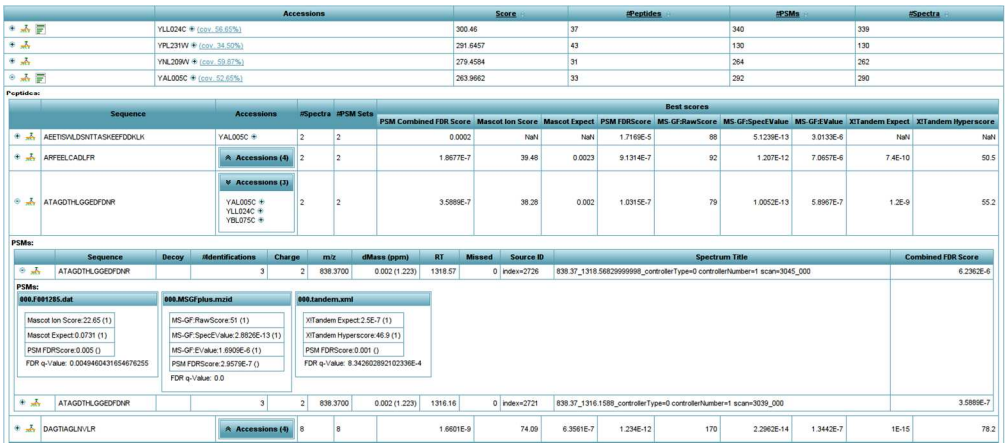


Figure 3. Screenshot of the inferred protein groups in the advanced viewer. For each protein group the accessions (with sequence coverage), the score and the number of peptides, PSMs and spectra are listed. Additionally the information on the peptides as well as the (combined) PSMs' information can be shown, which allows for an in-depth analysis of the inferred proteins.
173x75mm (300 x 300 DPI)

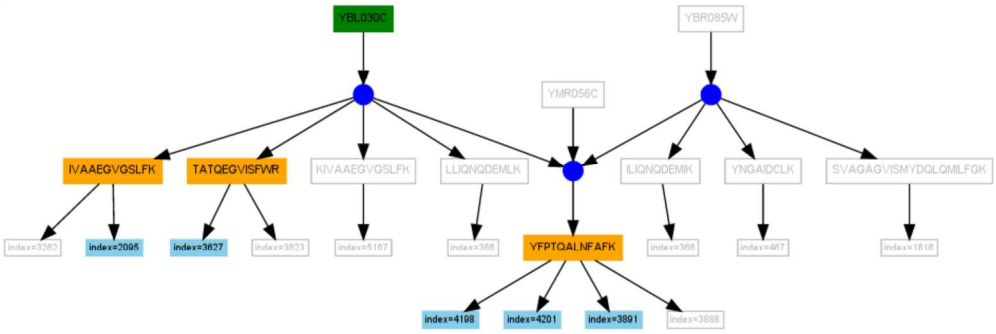


Figure 4. Visualization of the protein inference. The visualization of the PIA intermediate graph depicts, which spectra and corresponding peptides are assigned to which proteins after the inference. In this example, the highlighted spectra (light blue) and peptides (orange) are assigned to only one protein (green). The greyed out spectra and peptides are not assigned (due to FDR criteria) and the greyed out proteins are not reported due to inference settings (no FDR valid PSMs).
165x56mm (300 x 300 DPI)

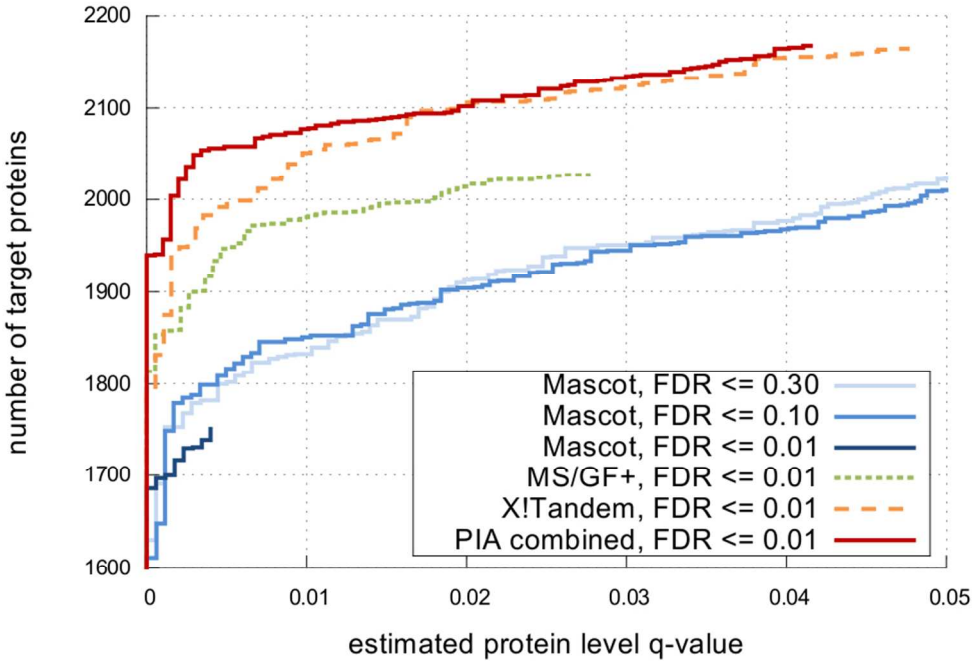


Figure 5. Performance of PIA on the Mouse Dataset. The figure plots in a pseudo ROC curve the number of target (in contrast to decoy) protein groups against the protein FDR q-values for protein inferences using the PSMs from three different search engines and the "Spectrum Extractor". The number of protein groups after a combination of search engine results with PIA exceeds the number of protein groups when using results of a single search engine at every q-value while using the same PSM FDR threshold. While decreasing the allowed FDR level also decreases the total number of reported proteins, the number of proteins in the low FDR range is increased, i.e. the beginning of the protein list contains fewer false positives. This increase of reported high quality proteins is only observable until a certain FDR level is reached, below which the number of reported proteins rapidly decreases, as plotted for the Mascot data.
84x57mm (300 x 300 DPI)

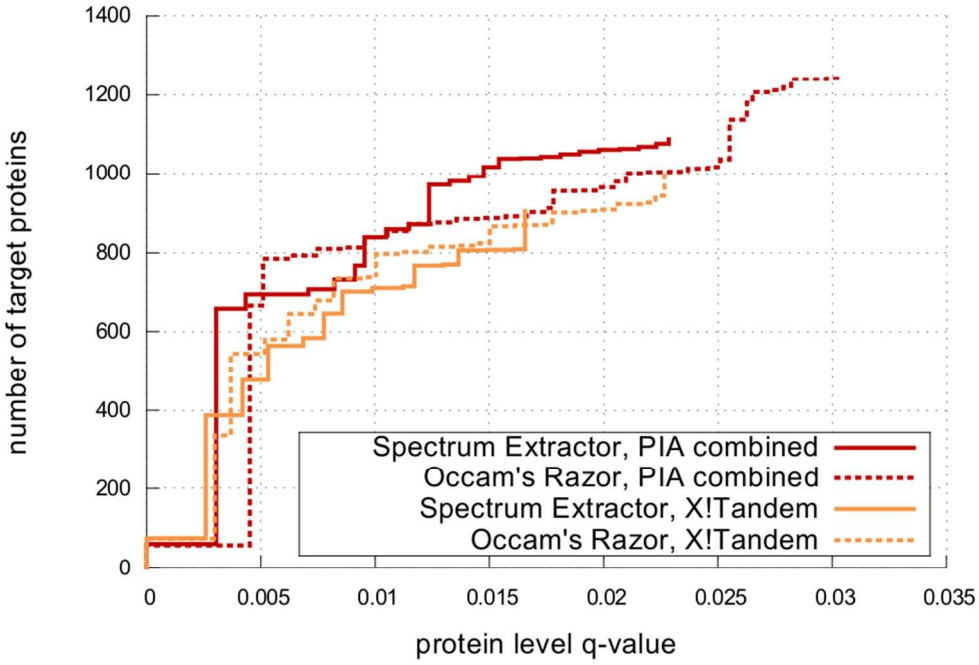


Figure 6. Performance of PIA on Yeast Gold Standard Dataset. For this dataset the expected identifications are known which allows the plotting of the number of true positive identifications against the q-value in a pseudo ROC curve. Plots are shown for protein inferences run with the "Spectrum Extractor" and "Occam's Razor" for a combination of search engines and the usage of X!Tandem results only. Generally the "Spectrum Extractor" outperforms the "Occam's Razor" in the very high confidence regions, but also tends to report fewer protein groups in the overall perspective.
84x57mm (300 x 300 DPI)

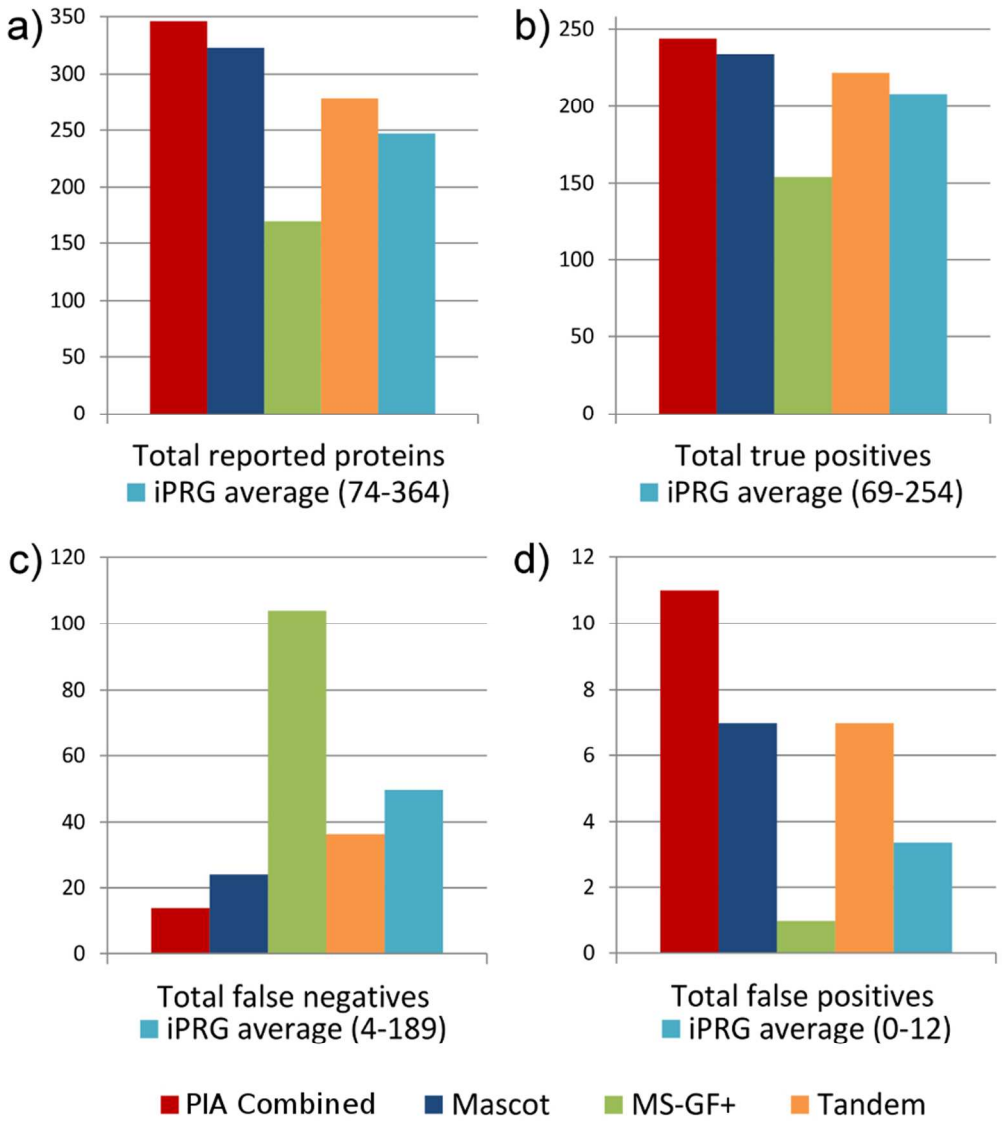


Figure 7. Performance of PIA on the iPRG 2008 dataset. (Caution: y-axes differ) The figure shows the number of (a) totally reported proteins, (b) the true positives, (c) the false negatives and (d) the false positives for the inferred proteins generated by PIA for either a combination of the search engine results or each search engine alone as well as the average result of the iPRG 2008 participants (in the parentheses the highest and lowest reported numbers are given). For the PIA analysis, the "Spectrum Extractor" with an FDR threshold of 0.01 was used on the PSM and protein level. It is seen, that PIA outperforms the average iPRG results, except when using the MS/GF+ results alone. For more details see text and supplemental.
83x96mm (300 x 300 DPI)