Journal of proteome research

Technical Note

pubs.acs.org/jpr

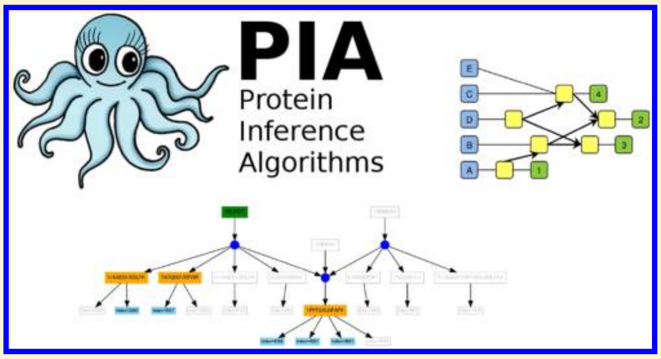
PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface

Julian Uszkoreit,[*] Alexandra Maerkens, Yasset Perez-Riverol,[‡] Helmut E. Meyer,[†,§] Katrin Marcus, Christian Stephan,[†,∥] Oliver Kohlbacher,[⊥] and Martin Eisenacher[*]

Medizinisches Proteom-Center, Ruhr-Universität Bochum, 44801 Bochum, Germany

Ⓢ *Supporting Information*

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 be run either 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 data sets, we show that PIA can identify a larger number of proteins at the same protein FDR when compared to that using 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 at https://github.com/mpc-bioinformatics/pia.



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

■ 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) are 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 are 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 of the same (sub)set of peptides, and it cannot be determined which of the proteins was 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 of these possible proteins (according to the inference algorithm) should be reported as suggested by Nesvizhskii et al.[9] and 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 the results cannot be completely explained or their ability to address a specific question cannot be determined. In addition 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]).

Received: February 6, 2015
Published: May 4, 2015

DOI: 10.1021/acs.jproteome.5b00121

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 a false discovery rate (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, the score in this context always means the score or probability). These scores are usually not directly comparable. Thus, they 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 FDR on the PSM and protein levels 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 be used either in a 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 performed 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 are based on three main concepts:

(1) A 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.

(2) A peptide, in contrast, refers to an amino acid sequence without charge state, either with or without regard to modifications, depending on user settings used for the inference.

(3) 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 that is stored in 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 to have 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, associated protein(s), charge state, precursor m/z value, difference between the measured and theoretical peptide masses, modifications, and 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 protein data.

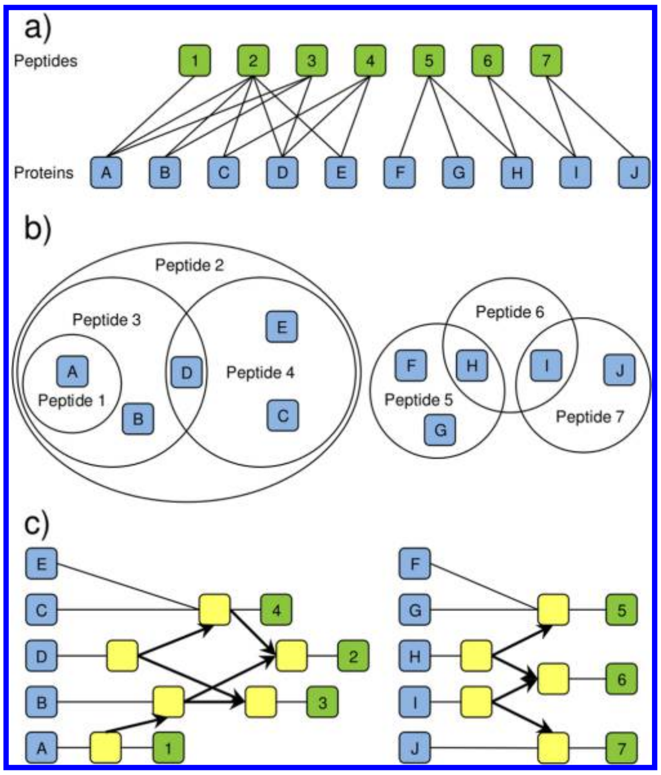


Figure 1. Compilation of the search engine results into a directed graph is performed in three steps. PSMs can be easily grouped into peptides according to their amino acid sequences; therefore, PSMs are left out in this figure. The connection information between peptides (green) and proteins (blue) 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 on 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 intermediate format used by PIA to quickly retrieve information. The yellow group nodes store no additional information, but they are necessary to connect the remaining nodes correctly and to uphold the set of rules given in the text (compilation of result files).

After collecting data from 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 evaluations (Figure 1a). 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 1b). These sets can be subsequently processed in parallel to consecutively insert each peptide into its corresponding acyclic graph compartment along with its protein accessions. The graph is constructed in a straightforward way and consists of nodes for proteins, peptides with their PSMs, and additional group nodes (Figure 1c). The group nodes connect the protein and peptide nodes such 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 the relations between proteins and peptides/PSMs to be retrieved rapidly.

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

Analysis of Identifications

This second step uses the compiled information on 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, then 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 to calculate the FDR, a search against a target-decoy database is recommended.[24,25] If such a search was conducted, then not only can a regular expression to distinguish decoy accessions from target accessions 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 as to whether modifications should be considered in order to distinguish peptides. This peptide step can be used to review the peptides and associated PSMs of proteins of interest or to obtain 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 filters' settings. Depending on the type of PSM scores (raw score, *p*-values, or *e*-values), different rules for 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, homologues, 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. However, 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 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, in addition to the choice of the actual inference algorithm. Currently, PIA implements three inference methods: (i) Report all, (ii) Occam's razor, and (iii) Spectrum Extractor.

(1) Report all: This is the simplest possible inference method, 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 that would be accepted in current publications, but it can be used to obtain a quick overview of the PSM and peptide data for a protein, which is actually not reported by any other method.
(2) 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 of the identified peptides that pass the given filters. Given the example in Figure 1 (and assuming no further filters), the protein groups with single proteins A, D, H, and I would be 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.
(3) 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 assigned to a peptide once, never gets assigned to another peptide. This concept is closer to reality, as, in most cases, one MS/MS spectrum contains only one peptide, although this may not always receive the highest score by 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, then 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. Afterward, the following steps are performed:

(1) For every protein group that has not yet been reported, examine each peptide. If a peptide is already reported, then allow it to be reported in this protein group with the prior set PSMs and score. Otherwise, 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, then use it for protein scoring only in the peptide where it has the best score.

Should there be more than one peptide in a protein group for which the spectrum has the best score, collect all spectra that may account for the affected peptides. If there are peptides that are in all of the affected spectra, then one of these peptides is used with all of the spectra while scoring, and 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 of its spectra assigned. If there are peptides with the same score and spectra, then all of their spectra are assigned, but only one is considered for 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, then assign it to the respective group appropriately. If it is not, then add the protein group and all of 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 of 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 the 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 subproteins are highlighted, as shown in Figure 4. The web interface is written for JavaServer Faces (JSF), which requires 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 PIA to be integrated into larger pipelines (e.g., using OpenMS).

Supported Data Formats

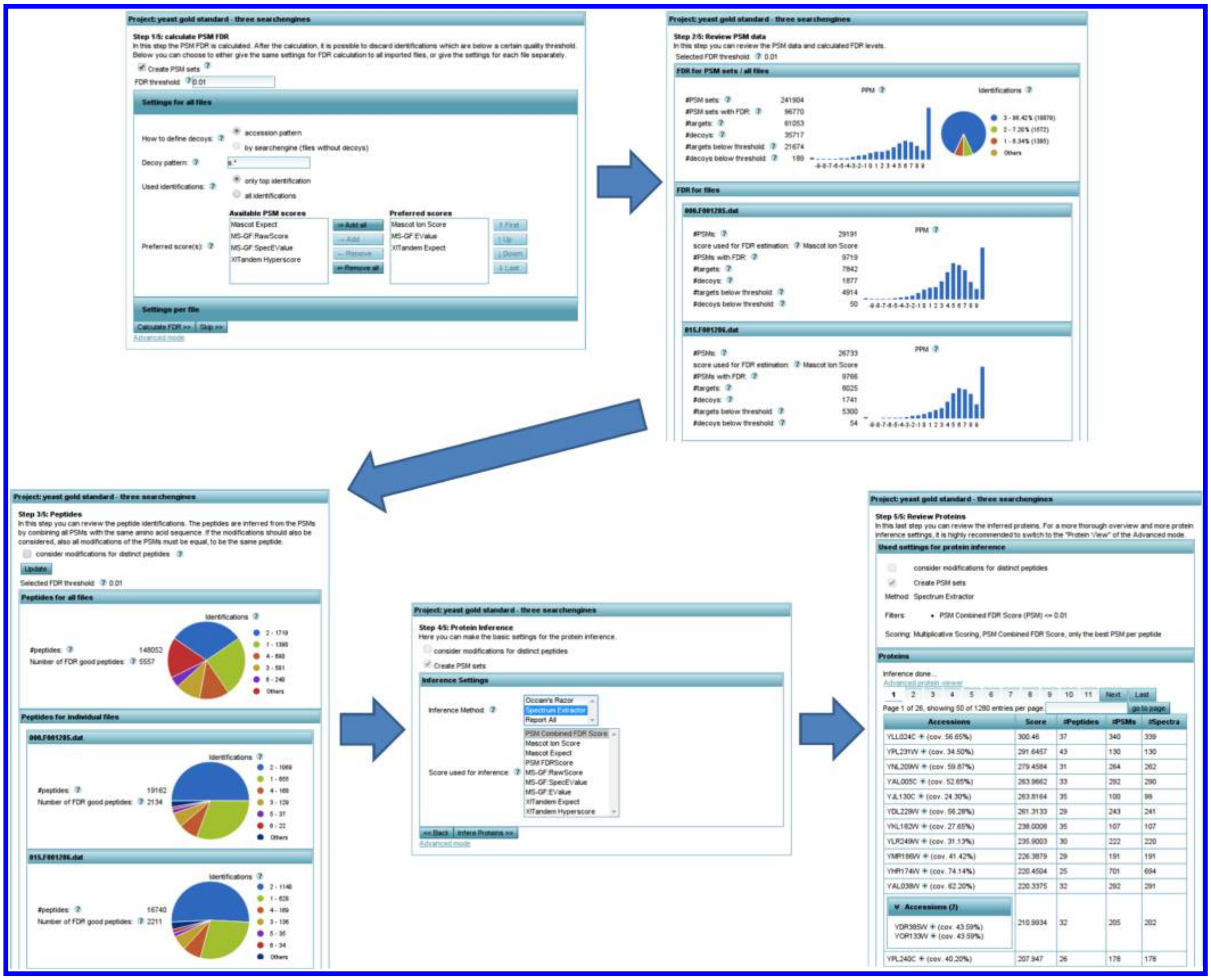
The current implementation supports importing mzIdentML[31] and thus import from virtually all search engines, provided a converter into this standard format exists already or the search engine directly exports into mzIdentML, like MS-GF+. Alternatively, the import of idXML files generated by OpenMS as well as the 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 has been 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 more easily human-readable and computer-parsable PSI format mzTab or a simple comma-separated values (CSV) format is also implemented.

Data Sets

To evaluate the reliability and to describe the behavior of PIA, it was assessed on one real-life in-house data set, of which the precise protein contents are not known, and on two public data sets with knowledge of the protein contents. The in-house data set is a label-free mass spectrometry analysis of a murine cell culture sample. The first data set with known protein content is part of the Gold Standard of Protein Expression in Yeast and was used by Ramakrishnan et al.[36] The other data set that also contains known proteins was produced for the Proteome Informatics Research Group (iPRG) 2008 study of the Association of Biomolecular Resource Facilities (ABRF). These data sets 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 data set and used to generate protein group results with different PIA settings and filters. The generation of the intermediate files and the report lists was performed on a laptop computer and took less than 1.5 h per data set.

All benchmarking data sets, the plotted search results, and used KNIME workflows have been deposited to the



Project: yeast gold standard - three searchengines

Step 1/5: calculate PSM FDR
In this step the PSM FDR is calculated. After the calculation, it is possible to discard identifications which are below a certain quality threshold. Below you can choose to either give the same settings for FDR calculation to all imported files, or give the settings for each file separately.

☑ Create PSM sets ?
FDR threshold: ? 0.01

Settings for all files

How to define decoys: ? ● accession pattern ○ by searchengine (files without decoys)
Decoy pattern: ? s.*
Used identifications: ? ● only top identification ○ all identifications

Preferred score(s): ?

Available PSM scores	Preferred scores
Mascot Expect	Mascot Ion Score
MS-GF:RawScore	MS-GF:EValue
MS-GF:SpecEValue	X!Tandem Expect
X!Tandem Hyperscore	

Settings per file

Project: yeast gold standard - three searchengines

Step 2/5: Review PSM data
In this step you can review the PSM data and calculated FDR levels.
Selected FDR threshold: ? 0.01

FDR for PSM sets / all files

#PSM sets: ?	241904
#PSM sets with FDR: ?	96770
#targets: ?	61053
#decoys: ?	35717
#targets below threshold: ?	21674
#decoys below threshold: ?	189

FDR for files

000.F001285.dat

#PSMs: ?	29191
score used for FDR estimation: ?	Mascot Ion Score
#PSMs with FDR: ?	9719
#targets: ?	7842
#decoys: ?	1877
#targets below threshold: ?	4914
#decoys below threshold: ?	50

015.F001286.dat

#PSMs: ?	26733
score used for FDR estimation: ?	Mascot Ion Score
#PSMs with FDR: ?	9766
#targets: ?	8025
#decoys: ?	1741
#targets below threshold: ?	5300
#decoys below threshold: ?	54

Project: yeast gold standard - three searchengines

Step 3/5: Peptides
In this step you can review the peptide identifications. The peptides are inferred from the PSMs by combining all PSMs with the same amino acid sequence. If the modifications should also be considered, also all modifications of the PSMs must be equal, to be the same peptide.

□ consider modifications for distinct peptides ?

Selected FDR threshold: ? 0.01

Peptides for all files

#peptides: ?	148052
Number of FDR good peptides: ?	5557

Peptides for individual files

000.F001285.dat

#peptides: ?	19162
Number of FDR good peptides: ?	2134

015.F001286.dat

#peptides: ?	16740
Number of FDR good peptides: ?	2211

Project: yeast gold standard - three searchengines

Step 4/5: Protein Inference
Here you can make the basic settings for the protein inference.

□ consider modifications for distinct peptides
☑ Create PSM sets

Inference Settings

Inference Method: ? Occam's Razor / Spectrum Extractor (selected) / Report All

Score used for inference: ? PSM Combined FDR Score / Mascot Ion Score / Mascot Expect / PSM FDRScore / MS-GF:RawScore / MS-GF:SpecEValue / MS-GF:EValue / X!Tandem Expect / X!Tandem Hyperscore

Project: yeast gold standard - three searchengines

Step 5/5: Review Proteins
In this last step you can review the inferred proteins. For a more thorough overview and more protein inference settings, it is highly recommended to switch to the "Protein View" of the Advanced mode.

Used settings for protein inference

□ consider modifications for distinct peptides
☑ Create PSM sets
Method: Spectrum Extractor
Filters:
- PSM Combined FDR Score (PSM) <= 0.01

Scoring: Multiplicative Scoring, PSM Combined FDR Score, only the best PSM per peptide

Proteins

Inference done...

Page 1 of 26, showing 50 of 1280 entries per page

Accessions	Score	#Peptides	#PSMs	#Spectra
YLL024C + (cov. 56.65%)	300.46	37	340	339
YPL231W + (cov. 34.50%)	291.6457	43	130	130
YNL209W + (cov. 59.87%)	279.4584	31	264	262
YAL005C + (cov. 52.65%)	263.9662	33	292	290
YJL130C + (cov. 24.30%)	263.8164	35	100	99
YDL229W + (cov. 56.28%)	261.3133	29	243	241
YKL182W + (cov. 27.65%)	238.0008	35	107	107
YLR249W + (cov. 31.13%)	235.9003	30	222	220
YMR186W + (cov. 41.42%)	226.3879	29	191	191
YHR174W + (cov. 74.14%)	220.4504	25	701	694
YAL038W + (cov. 62.20%)	220.3375	32	292	291
Accessions (2) YDR385W + (cov. 43.59%) YOR133W + (cov. 43.59%)	210.9934	32	205	202
YPL240C + (cov. 40.20%)	207.947	26	178	178

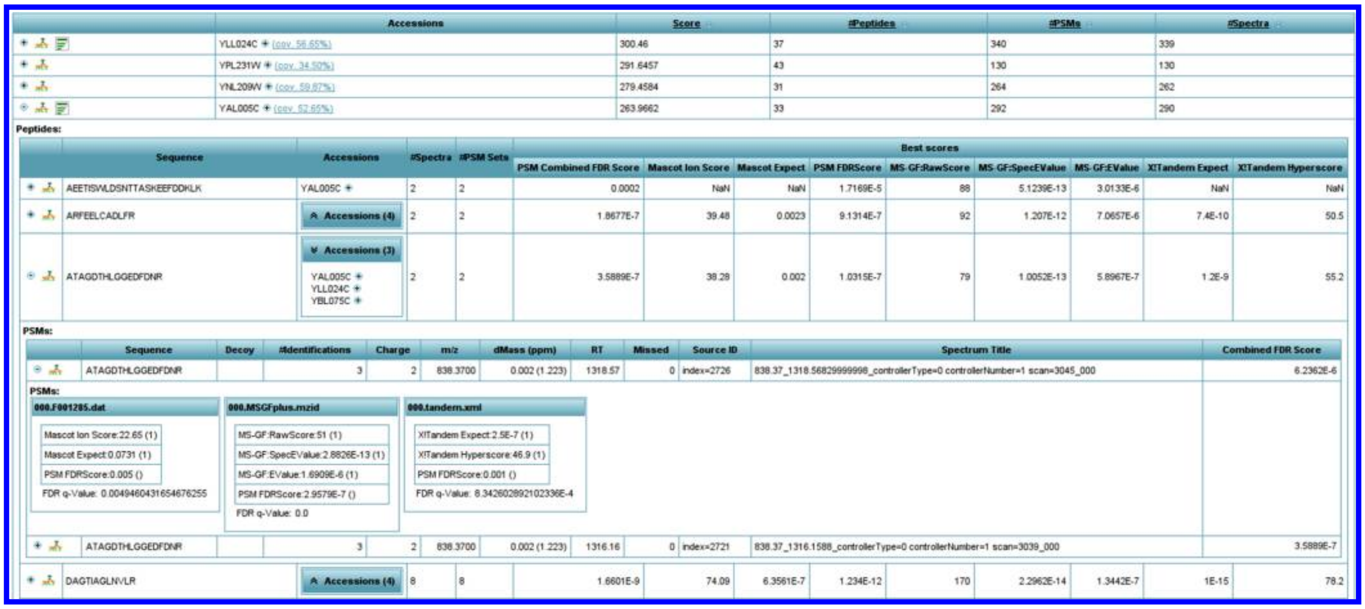
Figure 2. PIA web interface wizard. These screenshots demonstrate the steps of the PIA wizard, the most convenient way to perform an analysis. The wizard is part of the web interface and guides the user through the analysis while suggesting defaults for most of the used parameters. The wizard starts with the calculation of the PSM FDR values for each search engine run (step 1) and shows statistics on these values such as 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.

ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[37] with the data set identifiers PXD000790, PXD000792, and PXD000793.

Mouse Data Set

For the creation of the mouse data set, cultured cells of a murine myoblast cell line were harvested and centrifuged for 5 min at 800*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 s, on ice). After centrifugation (15 min, 16 000*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.[38] After digestion, the peptide concentration was determined by amino acid 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 Supporting Information 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 and 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, and carbamidomethylation of C was used as a fixed modification.



	Accessions	Score	#Peptides	#PSMs	#Spectra
+	YLL024C + (cov. 56.65%)	300.46	37	340	339
+	YPL231W + (cov. 34.50%)	291.6457	43	130	130
+	YNL209W + (cov. 59.87%)	279.4584	31	264	262
⊖	YAL005C + (cov. 52.65%)	263.9662	33	292	290

Peptides:

	Sequence	Accessions	#Spectra	#PSM Sets	Best scores								
					PSM Combined FDR Score	Mascot Ion Score	Mascot Expect	PSM FDRScore	MS-GF:RawScore	MS-GF:SpecEValue	MS-GF:EValue	X!Tandem Expect	X!Tandem Hyperscore
+	AEETISWLDSNTTASKEEFDDKLK	YAL005C +	2	2	0.0002	NaN	NaN	1.7169E-5	88	5.1239E-13	3.0133E-6	NaN	NaN
+	ARFEELCADLFR	∧ Accessions (4)	2	2	1.8677E-7	39.48	0.0023	9.1314E-7	92	1.207E-12	7.0657E-6	7.4E-10	50.5
⊖	ATAGDTHLGGEDFDNR	∨ Accessions (3) YAL005C + YLL024C + YBL075C +	2	2	3.5889E-7	38.28	0.002	1.0315E-7	79	1.0052E-13	5.8967E-7	1.2E-9	55.2

PSMs:

	Sequence	Decoy	#Identifications	Charge	m/z	dMass (ppm)	RT	Missed	Source ID	Spectrum Title	Combined FDR Score
⊖	ATAGDTHLGGEDFDNR		3	2	838.3700	0.002 (1.223)	1318.57	0	index=2726	838.37_1318.56829999998_controllerType=0 controllerNumber=1 scan=3045_000	6.2362E-6

PSMs:

000.F001285.dat

Mascot Ion Score:22.65 (1)

Mascot Expect:0.0731 (1)

PSM FDRScore:0.005 ()

FDR q-Value: 0.0049460431654676255

000.MSGFplus.mzid

MS-GF:RawScore:51 (1)

MS-GF:SpecEValue:2.8826E-13 (1)

MS-GF:EValue:1.6909E-6 (1)

PSM FDRScore:2.9579E-7 ()

FDR q-Value: 0.0

000.tandem.xml

X!Tandem Expect:2.5E-7 (1)

X!Tandem Hyperscore:46.9 (1)

PSM FDRScore:0.001 ()

FDR q-Value: 8.342602892102336E-4

	Sequence	Decoy	#Identifications	Charge	m/z	dMass (ppm)	RT	Missed	Source ID	Spectrum Title	Combined FDR Score
+	ATAGDTHLGGEDFDNR		3	2	838.3700	0.002 (1.223)	1316.16	0	index=2721	838.37_1316.1588_controllerType=0 controllerNumber=1 scan=3039_000	3.5889E-7

	Sequence	Accessions	#Spectra	#PSM Sets	PSM Combined FDR Score	Mascot Ion Score	Mascot Expect	PSM FDRScore	MS-GF:RawScore	MS-GF:SpecEValue	MS-GF:EValue	X!Tandem Expect	X!Tandem Hyperscore
+	DAGTIAGLNVLR	∧ Accessions (4)	8	8	1.6601E-9	74.09	6.3561E-7	1.234E-12	170	2.2962E-14	1.3442E-7	1E-15	78.2

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.

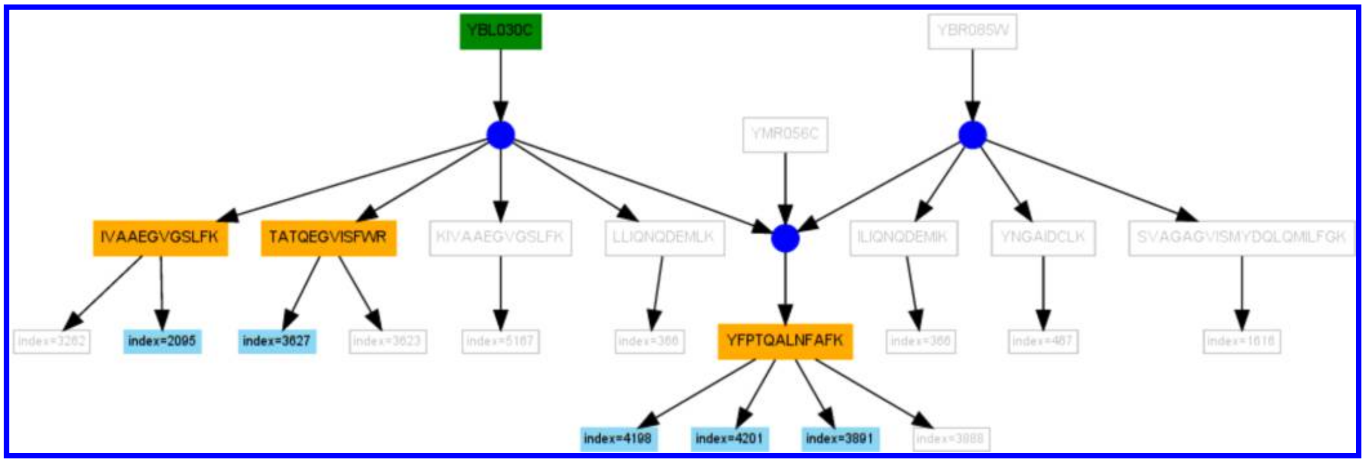


Figure 4. Visualization of the protein inference. Visualization of the PIA intermediate graph depicts which spectra and corresponding peptides are assigned to which proteins after the inference. In this example, the highlighted spectra (light blue) and peptides (orange) are assigned to only one protein (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).

Yeast Gold Standard Data Set

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 log phase. The expressed proteins contained in the sample were identified by MS- and non-MS-based methods and are available as a reference set. For the performance measurement on this data set, 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, 6717 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 known to be in the sample was adjusted (for more information, see Supporting Information file 1) and finally contains 4258 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 with the most spectra were used (070119-zl-mudpit07-1). For these runs, the RAW files were converted to mzML using 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 Data Set

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 were also provided for the ABRF iPRG208 study, which aimed to "assess the quality and consistency of protein reporting on a common data set", as stated on the study's slides. For this study, mouse samples were trypsin-digested, and peptides were 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 were measured multiple times with different exclusion lists), which resulted into 29 files. These data were 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 that 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 five 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 a 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 were used, and for the variable modifications, oxidation of M and protein N-terminal acetylation were used.

RESULTS AND DISCUSSION

Mouse Data Set

Using this data set, the application of PIA on a current data set was assessed. For this, the searches performed by Mascot, MS-GF+, and X!Tandem were first analyzed separately and then through a combination of all searches. The numbers of identified protein groups using 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 latter value is recommended by the authors), respectively. Although 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 observable only 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 number of proteins identified when using only MS-GF+, X!Tandem, and FDR below 0.01 are plotted, which show equal trends even though there are 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.

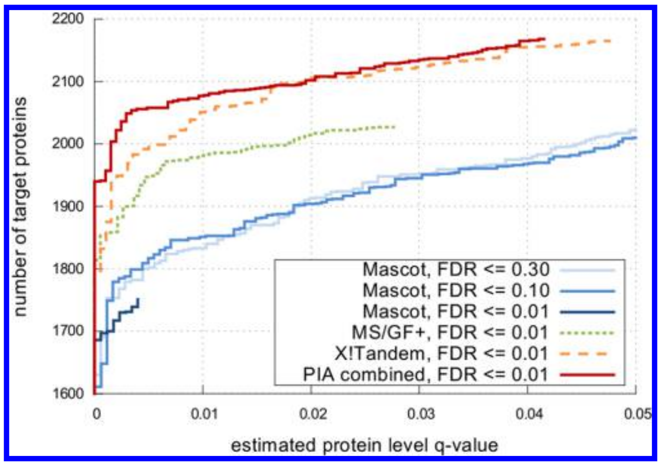


Figure 5. Performance of PIA on the mouse data set. Plotted is a pseudo ROC curve of the number of target (in contrast to decoy) protein groups against the protein FDR *q*-values for protein inferences using the PSMs from three different search engines and Spectrum Extractor. The number of protein groups determined after a combination of search engine results with PIA exceeds the number of protein groups identified when using results of a single search engine at every *q*-value while using the same PSM FDR threshold. Although 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 observable only until a certain FDR level is reached, below which the number of reported proteins rapidly decreases, as plotted for the Mascot data.

Yeast Gold Standard Data Set

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 data set. For this data set, the proteins contained in the sample are known, and 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 Supporting Information file 2).

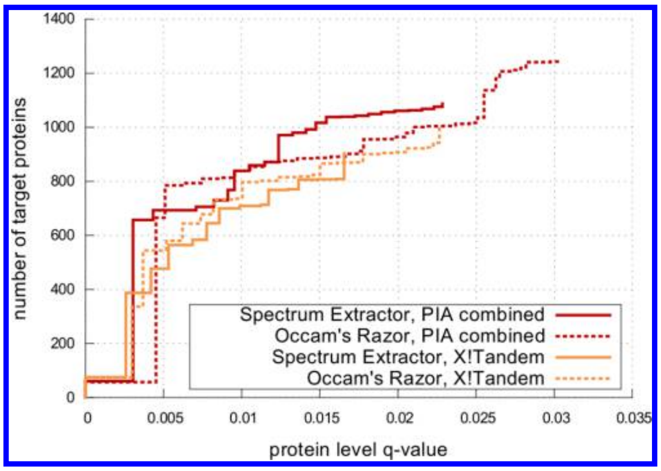


Figure 6. Performance of PIA on the yeast gold standard data set. For this data set, the expected identifications are known, which allows the number of true positive identifications to be plotted against the *q*-value in a pseudo ROC curve. Plots are shown for protein inferences run with Spectrum Extractor and Occam's Razor for a combination of search engines and the usage of X!Tandem results only. Generally, Spectrum Extractor outperforms Occam's Razor in the very high confidence regions, but it also tends to report fewer protein groups in the overall perspective.

Although the general behavior is similar, it is interesting to note that, for the data set used, 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. Although these observations are data set dependent, the data show overall good results for the inference algorithms used and do not make many false reports, as the plotted curves all stop before a value of 0.035. For all analyzed settings, the protein group with accession nos. 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 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 Data Set

In this data set, the expected number of identifications per cluster was calculated by the ABRF group members. For 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 a FDR below 0.01 for each search engine alone and in combination were inferred to a protein group list using Spectrum Extractor; for the comparison, only groups with a 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 interesting is that the number of true positives is much higher in the report from the combination, which is surpassed by only 6 iPRG participants (compare lists and charts in the sheet "overview2" of Supporting Information file 3). The false negative rates with the assessed PIA settings are better than the average iPRG participant's results. An exception is the MS/GF+ search, which reports the fewest proteins and therefore also has the highest number of false negatives, whereas the PIA combination is outperformed by only 6 of the iPRG participants. The relatively high number of false positives in all runs except the MS-GF+ run corresponds mainly to clusters, which are also dubious in the slides of the ABRF study (compare also Supporting Information 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, such as the need to have at least one unique peptide per protein, although stricter settings also decrease the total number of reported proteins and thus true positives.

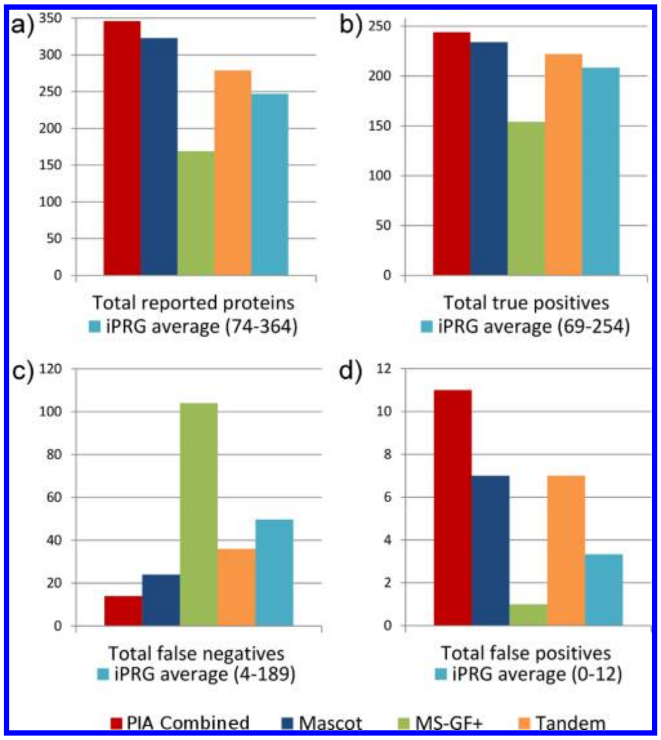


Figure 7. Performance of PIA on the iPRG 2008 data set. (note that the *y* axes differ). Number of (a) total reported proteins, (b) true positives, (c) false negatives, and (d) false positives for the inferred proteins generated by PIA for either a combination of the search engine results or each search engine alone as well as the average result of the iPRG 2008 participants (in parentheses are the highest and lowest reported numbers). For the PIA analysis, Spectrum Extractor with a FDR threshold of 0.01 was used on the PSM and protein levels. It can be seen that PIA outperforms the average iPRG results except when using the MS/GF+ results alone. For more details, see the text and Supporting Information.

For the data sets with known content, plots of the target-decoy estimated protein level *q*-values against the (claimed) true protein level *q*-values are shown in Supporting Information file 5. These plots show significant differences between the data sets. For the iPRG2008 data set, the estimated error is consistently much lower than the actual value, although the ratio goes down with the number of reported proteins. For the yeast gold standard data set, 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 data set, prior search results of the same actual MS data were used; thus, identifying more proteins than those that are claimed to be valid is more probable with different search engines. For determination of the yeast data set's content, other technologies were also used, which allows a more complete compilation of the proteins contained to be created. For an in-depth analysis of the estimation between the true and estimated protein *q*-values of protein inference algorithms, more data sets of complex protein mixtures having exactly known content would be needed, which are not available at the time of writing this manuscript.

■ CONCLUSIONS

In this article, 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 to improve the results of identification experiments. Different protein inference and scoring methods can be quickly tested, and their results be compared, either using a browser-based user-friendly interface that also allows for an in-depth analysis or using command line tools or KNIME nodes for pipeline environments.

DOI: 10.1021/acs.jproteome.5b00121

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 be easily inspected to find the peptides and even PSMs for any reported protein. The ability to export into easily parseable generic formats (CSV and mzTab) or more advanced formats like mzIdentML is included for further processing. With the ability to import files in mzIdentML format, we support virtually any search engine without a need for further implementation; in addition, native result files of some of the most used search engines can be imported, currently including 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 data sets 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 those used to perform fast and comparable spectral counting analyses, were used in recent publications[38,41,42] and prove the applicability of the results. Algorithms of PIA are included in a recent version of PRIDE Inspector[43] to allow protein inference and visualization using ms-data-core-api.[44] Future development tasks include implementing further protein inference and scoring algorithms to better discriminate isoforms, to make a decision regarding a protein group's representative accession, and to incorporate quantitation data.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed information about the generation of the SGD database (file 1), analysis of the yeast gold data set (file 2) and iPRG2008 data set (file 3), detailed description of the generation of the mouse data set (file 4), and comparison of estimated vs calculated q-values (file 5). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00121. PIA is released under a three-clause BSD license and freely available for download at https://github.com/mpc-bioinformatics/pia.

■ AUTHOR INFORMATION

Corresponding Authors

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

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

Present Addresses

[†](H.E.M. and C.S.) Medizinisches Proteom-Center, Ruhr-Universität Bochum, 44801 Bochum, Germany.

[‡](Y.P.-R.) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.

[§](H.E.M.) Leibniz-Institut für Analytische Wissenschaften − ISAS − e.V. ,44139 Dortmund, Germany

[∥](C.S.) Kairos GmbH, 44799 Bochum, Germany.

[⊥](O.K.) Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center, and Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.U. and M.E. are funded by PURE (Protein Unit for Research in Europe), a project of North Rhine-Westphalia, Germany, and by the BMBF grant de.NBI - German Network for Bioinformatics Infrastructure (FKZ 031 A 534A). A.M. is funded by the German Research Foundation (FOR1228). Y.P.-R. is supported by the BBSRC "PROCESS" grant [reference BB/K01997X/1]. 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*, 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*, 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. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.

(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*, 3551−67.

(5) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 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*, 2840−52.

(7) Bandeira, N. Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques* **2007**, *42*, 687−9.

(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*, 3736−47.

(9) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics* **2005**, *4*, 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*, 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*, M111.014381.

(12) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Briefings Bioinf.* **2012**, *13*, 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*, 4646−58.

(14) Searle, B. C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010**, *10*, 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*, 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*, 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*, 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*, 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**, *13*, 2765−75.

(21) Orchard, S.; Hermjakob, H.; Apweiler, R. The proteomics standards initiative. *Proteomics* **2003**, *3*, 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 Anal., Mach. Learn. Appl.* **2008**, 319−26.

(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 Bioinf.* **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*, 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*, 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. *Mol. Cell. Proteomics* **2006**, *5*, S216−S216.

(27) *Apache Tomcat*; Apache Software Foundation: Los Angeles, CA. 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*, 5th International Workshop on Science Gateways, Zurich, Switzerland, June 3−5, 2013.

(30) *GenericKnimeNodes*; GitHub. https://github.com/genericworkflownodes/GenericKnimeNodes.

(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*, 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*, 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*, 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*, 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**, *14*, 2389−99.

(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*, 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*, 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*, 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*, 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*, 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*, 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**, *90*, 14−27.

(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*, 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**, DOI: 10.1093/bioinformatics/btv250.