A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms

Matthew The[†], **Fredrik Edfors**[†], **Yasset Perez-Riverol**[‡], **Samuel H. Payne**[§], **Michael R. Hoopmann**[∥], **Magnus Palmblad**[⊥], **Björn Forsström**[†], and **Lukas Käll**[*,†]

[†]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH — Royal Institute of Technology, Box 1031, 17121 Solna, Sweden [‡]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom [§]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington United States [∥]Institute for Systems Biology, Seattle, Washington 98109, United States [⊥]Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands

Abstract

A natural way to benchmark the performance of an analytical experimental setup is to use samples of known composition and see to what degree one can correctly infer the content of such a sample from the data. For shotgun proteomics, one of the inherent problems of interpreting data is that the measured analytes are peptides and not the actual proteins themselves. As some proteins share proteolytic peptides, there might be more than one possible causative set of proteins resulting in a given set of peptides and there is a need for mechanisms that infer proteins from lists of detected peptides. A weakness of commercially available samples of known content is that they consist of proteins that are deliberately selected for producing tryptic peptides that are unique to a single protein. Unfortunately, such samples do not expose any complications in protein inference. Hence, for a realistic benchmark of protein inference procedures, there is a need for samples of known content where the present proteins share peptides with known absent proteins. Here, we present such a standard, that is based on *E. coli* expressed human protein fragments. To illustrate the application of this standard, we benchmark a set of different protein inference procedures on the data. We observe that inference procedures excluding shared peptides provide more accurate estimates of errors compared to methods that include information from shared peptides, while still giving a reasonable performance in terms of the number of identified proteins. We also demonstrate that using a sample of known protein content without proteins with shared tryptic peptides can give a false sense of accuracy for many protein inference methods.

[*]**Corresponding Author:** lukas.kall@scilifelab.se.

The authors declare no competing financial interest.

INTRODUCTION

Shotgun proteomics offers a straightforward method to analyze the protein content of any biological sample. The method involves proteolytic digestion of the proteins into peptides, which greatly improves the efficiency of the technique but also introduces a problem for the subsequent data processing. As the mass spectrometers are detecting ions from peptides rather than proteins directly, the evidence for the detection of the peptides has to be integrated into evidence of the presence of proteins in the original sample, using a *protein inference* algorithm.[1]

This protein inference procedure is complicated by the homology within most proteomes; many proteins share constituent proteolytic peptides, and it is not clear how to best account for such *shared peptides*. For example, should we see shared peptides as evidence for all, a subset, or none of its potential aggregate proteins? While the field of computational proteomics starts to reach a consensus on how to estimate the confidence of peptide-spectrum matches (PSMs) and peptides, there is still relatively little work done in establishing standards that evaluate how much confidence we can give reported protein inferences or even what the best methods to infer proteins from shotgun proteomics data are.

Currently, there are two available methods to determine the accuracy of inference procedures and their error estimates: (i) simulations of proteomics experiments and (ii) analysis of experiments on samples with known protein content. By simulating the proteolytic digestion and the subsequent matching of mass spectra to peptides,[2–4] one can obtain direct insights into how well the simulated absence or presence of a protein is reflected by a protein inference procedure. However, there is always the risk that the assumptions of the simulations are diverging from the complex nature of a mass spectrometry experiment. For instance, there is currently a lack of accurate tools to predict which peptides typically would be detected in an experiment or how well their spectra will be matched by a typical search engine. Hence, accurate predictions on simulated data can only be viewed as a minimum requirement for a method to be considered accurate.[4]

A more direct characterization of protein inference procedures can be obtained by analyzing experiments on samples of known protein content.[5] For such experiments, a protein standard is assembled from a set of isolated and characterized proteins and subsequently analyzed using shotgun proteomics. Normally, the acquired spectra are searched against a database with sequences from proteins known to be present, as well as absent proteins.[6] Two notable protein standards are available today, the ISB18[5] and the Sigma UPS1/UPS2 standard. Both standards consist of a relatively limited set of proteins, 18 proteins in the ISB18 and 48 proteins in the UPS. However, neither of those standards produce tryptic peptides shared between multiple protein sequences. Hence, these standards are a poor fit for benchmarking

protein inference algorithms, as the real difficulty of protein inferences is proteolytic peptides shared between multiple proteins.

Here, we present a data set specifically designed for benchmarking proteoform inference algorithms. Two different samples of defined content were created from pairs of proteins that share peptides. These protein standards are a byproduct from the antibody production from the Human Proteome Atlas project (http://www.proteinatlas.org/),[7] where protein fragments, referred to as *Protein Epitope Signature Tags* (PrESTs), are expressed in recombinant *E. coli* strains as antigens to be injected into rabbits to raise polyclonal antibodies. We demonstrate that such protein fragments can be used for benchmark protein inference procedures and for evaluating the accuracy of any confidence estimates such methods produce. This data set has also previously been made available in anonymized form as a part of the iPRG2016 study (http://iprg2016.org) inference. Here, we also evaluated a set of principles for proteoform inference with the sets.

METHODS

Data Generation

To generate the data sets, recombinant protein fragments (PrESTs) from the Human Proteome Atlas project were scanned for a total of 191 overlapping PrEST sequences after in-silico trypsin digestion. PrEST fragments were cloned into the expression pAff8c and thereafter transformed into an *E. coli* strain for recombinant protein production following verification by DNA sequencing. Cells containing expression vectors were cultivated as previously described.[8] Cell cultures were harvested and the QPrESTs were purified using the N-terminal quantification tag that included a hexahistidine tag used for immobilize metal ion affinity chromatography (IMAC) and stored in 1 M urea, 1xPBS. Each protein fragment is thereafter individually evaluated by SDS-PAGE analysis, and the molecular weight is verified by LC-MS analysis. The total library of PrESTs produced within the Human Protein Atlas program was used for selection, and all available pairs were included in this study. All PrESTs are fused together with a solubility tag (Albumin Binding Protein, 163 aa), and all PrEST sequences are of a similar length ranging between 50 and 150 aa.

To generate the data sets, the PrEST sequences of the Human Proteome Atlas-project were scanned for a total of 191 overlapping pairs of PrEST sequences. The resulting protein pairs share at least one identical fully tryptic peptide. Each pair originates from different human genes, and flanking amino acids around the tryptic termini are thereby not shared between them. Each recombinant protein was randomly assigned into one of the pools (A or B) that were created (Figure 1).

A third pool was created by mixing pools A and B, resulting in a pool A+B. An amount of 1.8 pmol of each PrEST was added to either pool A or pool B, and an amount of 0.9 pmol of all PrESTs was added to pool A+B. The total protein amount of 10 μg from each of the pools was reduced with dithiothreitol and alkylated with iodoacetamide prior to trypsin digestion overnight, and each pool was mixed into a background of a tryptic digest of 100 ng *Escherichia coli* [BL21(DE3) strain], resulting in three mixtures, Mixture A, mixture B, and mixture A+B. In effect, the resulting concentrations of tryptic peptides shared between

mixture A and mixture B were the same across all three mixtures, while peptides unique to mixture A or mixture B appear in half the concentration in mixture A+B.

The E. coli digest was generated from a nontransformed bacterial strain cultivated and harvested alongside the PrEST producing strain. Cells were dissolved in Lysis buffer (100 mM Tris-HCl, 4% SDS, 10 mM DTT, pH 7.6) and incubated at 95 °C in a thermomixer for 5 min at 600 rpm and thereafter sonicated at 50% amp (1s pulse, 1 s hold) for 1 min. The lysate was diluted with denaturing buffer (8 M urea, 100 mM Tris- HCl, pH 8.5) and centrifuged through a 0.22 μm spin filter (Corning, Corning, NY, USA). Trypsin digestion was performed using a previously described filter-aided sample preparation (FASP) method. [9]

A per sample amount of 1.1 μg of each of three mixtures was analyzed in triplicate by LC-MS/MS in random order. The digests were loaded onto an Acclaim PepMap 100 trap column (75 μm × 2 cm, C18, 3 μm, 100 Å), washed for 5 min at 0.25 $\mu L/$ min with mobile phase A [95% H_2O, 5% DMSO, 0.1% formic acid (FA)], and thereafter separated using a PepMap 803 C18 column (50 cm × 75 μm, 2 μm, 100 Å) directly connected to a Thermo Scientific Q-Exactive HF mass spectrometer. The gradient went from 3% mobile phase B [90% acetonitrile (ACN), 5% H_2O, 5% DMSO, 0.1% FA] to 8% B in 3 min, followed by an increase up to 30% B in 78 min and thereafter an increase to 43% B in 10 min followed by a steep increase to 99% B in 7 min at a flow rate of 0.25 μL/min. Data were acquired in data-dependent (DDA) mode, with each MS survey scan followed by five MS/ MS HCD scans (AGC target 3e6, max fill time 150 ms, mass window of 1.2 m/z units, the normalized collision energy setting stepped from 30 to 24 to 18 regardless of charge state), with 30 s dynamic exclusion. Charge states of 2–7 were selected for fragmentation, while charge state 1 and undefined charge states were excluded. Both MS and MS/MS were acquired in profile mode in the Orbitrap, with a resolution of 60,000 for MS and 30,000 for MS/MS.

Each replicate injection of each mixture (rep 1–3) was separated by one wash followed by a blank injection (*E. coli* digest) in order to limit potential cross contamination between mixtures. Injections were made in the following order, blank first replicate, Mixture B (X3), blank second replicate, Mixture A (X3), blank third replicate, and Mixture A+B (×3).

Data Set

We assembled the data into a test data set consisting of the following:

- Three FASTA-files, containing the amino acid sequence of the Protein fragments of mixture A, mixture B. We also include a set of *entrapment* sequences of 1000 representative PrEST fragments that were not selected for the mixtures and hence were absent from the samples.

- These entrapment sequences had a similar 50–150 amino acid long length distribution as the PrEST sequences of our mixtures.

- Twelve runs, consisting of triplicates of analyses of mixture A, mixture B, mixture A+B, and "blank" runs without spike-ins, all run in a background of *E.*

coli-lysates. These are provided in Thermo raw data-format. The number of MS2 spectra per run ranges from 27363 to 28924.

- •   An evaluation script, written in python.

The FASTA files and scripts can be downloaded from https://github.com/statisticalbiotechnology/proteoform-standard, and the mass spectrometry data is accessible from the pride database under the project accession number PXD008425.

This data set has been made available in anonymized form as a part of the iPRG2016 study (http://iprg2016.org) on protein inference. The sample mixtures can be made available on request, for evaluation under other mass spectrometers than the one we used in this study.

Data Processing

The raw data files were converted to MS1 and MS2 format using ProteoWizard[10] and subsequently processed by Hardklor[11] followed by Bullseye,[12] through the interface of the Crux 2.1 package,[13] to set monoisotopic masses. The resulting ms2 spectra were then matched to separate target and decoy sequence databases (described below) using Tide within Crux 2.1[13,14] (no variable modifications, 10 ppm search tolerance, all other parameters kept at their default values) and Percolator v3.01,[15,16] deriving peptide-level probabilities for each of the mass spectrometry runs.

The target protein database consisted of all three FASTA files combined, and the decoy database was constructed by reversing the protein sequences of the target database.

For each of the runs we calculated protein-level *Entrapment FDRs*[6] by counting all matches to PrESTs present in their analyzed mixture as correctly matched, and all matches to PrESTs absent (i.e., stemming from the set of 1000 nonpresent PrESTs or from the set of the mixtures not used in the sample) as being incorrectly matched. As the correct matches only map to present proteins, whereas incorrect matches distribute over both present and absent proteins, we also normalized the Entrapment FDR by the prior probability of the PrEST to be absent, the so-called π_A.[4]

RESULTS

We constructed two mixtures, each containing 191 PrESTs, that is one out of each of the 191 pairs of PrEST sequences with partially overlapping amino acid sequence (Figure 1). We analyzed the two samples as well as a combination of the two using LC-MS/MS (see the Methods section).

The three data sets make an informative benchmark set. By matching the spectra of the data set against a bipartite database containing both the present and some absent PrEST sequences, we obtain a direct way to count the number of inferred PSMs, peptides, and proteins stemming from nonpresent PrESTs.[6] More specifically this allows us to assess the fraction of identifications in a set that stems from absent PrESTs, which we here will refer to as the *Entrapment FDR.* However, unlike traditional samples of known content, this standard

contains overlapping protein fragments, which allows us to assess the performance of protein inference algorithms in the presence of homology.

Methods To Infer Protein Sequences

We set out to test a set of different protein inference algorithms against our test set. We first analyzed the data using Crux[13,14] and Percolator,[15,16] deriving peptide-level probabilities for each combination of triplicate mass spectrometry runs. We subsequently compared the performance of the different schemes for inferring proteins and their confidence.

First, there are different ways to infer proteins from peptide sequences. The major difference between the methods relates to how they handle so-called *shared peptides,* that is peptide sequences that could stem from more than one protein. The tested inference methods were the following:

Inclusion.—Possibly the easiest way to handle shared peptides is to assign any found peptide to all its possible causative proteins. Under this assumption, we infer the presence of any protein which links to an identified peptide.

Exclusion.—Another method is to remove any shared peptides before any reconstruction takes place. Under this assumption, we infer the presence of any protein which links to an identified peptide unique to the protein.

Parsimony.—A method that is quite popular for handling shared peptides is to use the principle of parsimony, i.e. to find the minimal set of proteins that would best explain the observations of the PSMs with a score above a given threshold. This principle has been implemented in a couple of well-known software tools such as IDPicker 2.0[17] and MaxQuant.[18] In cases where there are multiple such minimal sets of proteins, several strategies can be used: do not include any of the sets, apply some form of protein grouping (see the Discussion section), or select one of the sets, either at random or based on the order the proteins are listed in the database. Here, we have opted to use the latter alternative, to select one of the sets at random.

Here, it is worth noting that our benchmark did not include methods aimed at inferring protein-groups, as we want to compare the methods ability to infer proteoforms.

Methods To Rank Protein Identifications

More than just inferring the protein sequences, any practically usable protein inference strategy has to assign confidence estimates in terms of posterior probabilities or false discovery rates. One way to assign such protein-level statistics is by investigating decoy ratios, a process that depends on assigning scores to rank our confidence in the different protein sequences. The different methods to obtain protein-level scores differ in the way they combine the confidence estimates of the proteins' constituent peptides. We tested five different methods to score proteins:

Products of PEPs.—This method summarizes a score for the protein's constituent peptides by calculating the product of peptide-level posterior error probabilities (PEPs).[18]

This method has been extensively used by tools such as MaxQuant,[18] PIA,[19] and IDPicker.[17]

Fisher's Method.—Another method that relies on an assumption of independence between different peptides' incorrect assignments to a protein is Fisher's method for combining independent *p* values, which is a classical technique for combining *p* values.[20] Fisher's method takes into account all constituent peptides of a protein,[21–23] by summarizing the individual peptides empirical *p* values. Unlike the product of PEPs, which also combines peptide-level evidence, Fisher's method explicitly accounts for the number of *p* values being combined and hence normalizes for protein length to some extent.

Best Peptide.—Instead of weighting together peptide-level evidence for a protein, some investigators chose to just use the best available evidence for a protein.[24] Savitski et al.[25] showed that, on large-scale data sets, taking the best-scoring peptide as the representative of a protein was superior to incorporating information from lower-scoring peptides. This approach might feel unsatisfying for most investigators, as the method discards all information but the best-scoring PSM for each protein.

Two peptides.—A simple way to combine evidence at the peptide level is the widely used two-peptide rule.[26] This approach requires evidence for a second peptide to support a protein inference, thereby preventing so-called "one-hit wonders", i.e., cases where a single, potentially spurious PSM yields a spurious protein detection.[24] Furthermore, the recently published Human Proteome Project Guidelines for Mass Spectrometry Data Interpretation version 2.1 requires "two non-nested, uniquely mapping (proteotypic) peptides of at least 9 aa in length", to count a protein sequence as being validated with mass spectrometry.[27]

Fido.—A more elaborate method to estimate the confidence in inferred proteins is to use Bayesian methods, represented here by Fido,[28] which calculates posterior probabilities of proteins' presence or absence status given the probabilities of peptides correctly being identified from the mass spectrometry data. Such methods are normally seen as inference procedures, but due to the design of our study, we listed it as a confidence estimation procedure to be used in combination with Inclusion inferences. Fido's use requires selection of prior probabilities for protein presence, present proteins emitting detectable peptides, and mismatches. These probabilities were set using grid searches, as implemented when running Fido through Percolator.[15]

In Supporting Information Table S-4 we have mapped a set of commonly used protein inference tools according to the names of the inference and ranking principles we use in this paper.

Benchmark

We wanted to distinguish the measurement of the performance of the inference methods, i.e. the ability of the methods to accurately rank protein inferences, from the accuracy of the methods, i.e. the calibration of the methods reported statistics. We hence have divided the benchmark of the tested inference methods into two subsections below.

Performance

First, we measured the performance of the permutations of inference and confidence estimation procedures, in terms of the number of identified proteins at a 5% protein-level entrapment- FDR (see Table 1 and S-1, as well as Figure S-1). It is worth noting that there is a fundamental difference between the different benchmark sets,; for the set A+B all proteins that share tryptic peptides are present. However, the sets A and B contain tryptic peptides shared between absent and present proteins.

Comparing the different methods of dealing with shared peptides, we found that the Inclusion and Parsimony methods reported more proteins than the Exclusion method when investigating the A+B set that contains peptides shared between present proteins. However, for the A set that contains peptides shared between the present and absent proteins, the Exclusion reports more proteins than the Inclusion methods (except Fido) and the Parsimony methods.

For the different confidence estimation procedures, we noted that the TWO PEPTIDES method reported fewer proteins than the other methods, whereas FIDO reports more proteins than the other methods.

Many implementations of Parsimony are two-step procedures, which first threshold on peptide- or PSM-level FDR and subsequently infer the most parsimonious set of proteins. In such implementations, one ends up controlling the list of proteins at both peptide and protein-level. We chose to make a more extensive test series of Parsimony for 1%, 5%, and 10% peptide- level FDR (see Figure S-3, Table S-2, and Table S-3). A trend is observable for such data; for the A+B set, more proteins are observed for similar protein-level FDRs when using a higher peptide-level FDR (e.g., 10%) than when using a lower peptide- level FDR (e.g., 1%). However, the inverse is true for the A set; more proteins are observed for similar protein-level FDRs when using a 1% peptide-level FDR than when using a 10% peptide-level FDR. This is a consequence of the cases where the algorithm has to select one of many minimal sized subsets explaining the observed peptides at random. Such a selection is not harmful to the performance if all minimal subsets consist of present proteins, but it might be harmful if one or more of the minimal subsets contains absent proteins.[29]

Accuracy of Confidence Estimates

Subsequently, we set out to estimate the accuracy of the different confidence estimates of the protein inference procedures and hence plotted the entrapment FDR as a function of the methods' reported FDR in Figure 2 and Supplementary Figure S-2.

All the tested methods reported acceptable and similar accurate statistics for the sets where the shared peptides stem from proteins that are all present (set A+B). However, when comparing the decoy and entrapment FDR for the sets with peptides shared between present and absent proteins (set A) we see that none of the inference methods using Inclusion or Parsimony reported satisfying decoy-derived FDRs.

The Exclusion principle, on the other hand, seemed to handle the sets with peptides shared between present and absent proteins in a satisfying manner, and particularly the Fisher's

method and the Products of PEPs gave quite accurate statistics for such data without reducing the number of reported proteins.

DISCUSSION

Here we have described a protein standard that can be used for comparing different algorithmic approaches to inferring sets of proteins from shotgun proteomics data. The standard is particularly useful for determining how to handle protein inferences in cases where a peptide could stem from multiple different proteoforms. We used the data set to compare a set of approaches for protein inference and protein scoring models, and we found that the reliability of protein inferences became more accurate when excluding any peptides shared between multiple proteins as compared to, e.g., inferring the most parsimonious set of proteins.

Many algorithms and tools that use Parsimony or Inclusion principles group proteins according to the peptide evidence that they share. Particularly in the case of Inclusion, the accuracy would increase dramatically if we would have evaluated protein groups rather than individual proteins. We have not included this option here for the sake of simplicity, not least because it confounds the null hypothesis in a way that complicates a fair comparison between methods.[29,4]

While the set is larger and more complex than other samples of known content (several millions of spectra), we see room for future improvements in terms of both more and longer protein sequences than our current standard as well as more complicated patterns of shared peptides.

Recently, a couple of other benchmarks for protein inference algorithms have been published. First, when selecting a protein inference strategy for Percolator, the authors used simulations to show that excluding shared peptides and scoring the protein based on the best scoring peptide performed overall better than the compared methods. However, on small-sized data sets, the method of multiplying PEPs had a slight performance advantage.[16]

Second, a large center study, Audain et al.,[30] benchmarked a set of protein inference algorithm implementations on "gold standards", i.e. manually annotated data sets. The authors conclude that PIA[19] and Fido[28] perform better than the other analyzed implementations on their data sets. We did not include PIA in our comparisons, but we did include Fido. In line with Audain et al., Fido gave an excellent performance and calibration and on data sets where all proteins sharing a peptide were present (set A+B). However, Fido's assessment of reliability scores was less than ideal for the data sets where the shared peptides' causative proteins were from both the present and absent groups (set A and set B). This behavior could not have been characterized using data sets lacking peptides shared between multiple proteins, and it is hence not a surprise that such characteristics have not been noted in previous studies.

The two-peptide rule performed poorly, as has been reported by several studies.[24,31,32] For instance, Veenstra et al. wrote already in 2004 that "Simply disregarding every protein identified by a single peptide is not warranted".[24] Similarly, the FDRs reported after using

parsimony to handle shared peptides seems, in general, to be anticonservative which is in line with what was reported by Serang et al.: "parsimony and protein grouping may actually lower the reproducibility and interpretability of protein identifications."[29]

In this study we have kept the peptide inference pipeline identical for all protein inference pipelines, to enable a *ceteris paribus* comparison of the protein inference methods. We also have tried to benchmark principles rather than implementations but made an exception for the Fido inference method, as this was readily available in the Percolator package that was used for the peptide inferences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

REFERENCES

(1). Nesvizhskii AI; Aebersold R Interpretation of shotgun proteomic data the protein inference problem. Mol. Cell. Proteomics 2005, 4, 1419–1440. [PubMed: 16009968]

(2). Bielow C; Aiche S; Andreotti S; Reinert K MSSimulator: Simulation of mass spectrometry data. J. Proteome Res 2011, 10, 2922–2929. [PubMed: 21526843]

(3). Serang O; Käll L Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. J. Proteome Res 2015, 14, 4099–4103. [PubMed: 26257019]

(4). The M; Tasnim A; Kall L How to talk about protein-level False Discovery Rates in shotgun proteomics. Proteomics 2016, 16, 2461–2469. [PubMed: 27503675]

(5). Klimek J; Eddes JS; Hohmann L; Jackson J; Peterson A; Letarte S; Gafken PR; Katz JE; Mallick P; Lee H; Schmidt A; Ossola R; Eng JK; Aebersold R; Martin DB The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. J. Proteome Res 2008, 7, 96–103. [PubMed: 17711323]

(6). Granholm V; Noble WS; Kall L On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J. Proteome Res 2011, 10,2671–2678. [PubMed: 21391616]

(7). Uhlen M; Oksvold P; Fagerberg L; Lundberg E; Jonasson K; Forsberg M; Zwahlen M; Kampf C; Wester K; Hober S; Wernerus H; Björling L; Ponten F Towards a knowledge-based human protéin atlas. Nat. Biotechnol 2010, 28, 1248–1250. [PubMed: 21139605]

(8). Nilsson P; Paavilainen L; Larsson K; Ödling J; Sundberg M; Andersson A-C; Kampf C; Persson A; Szigyarto CA-K; Ottosson J; Björling E; Hober S; Wernérus H; Wester K; Pontén F; Uhlén M Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. Proteomics 2005, 5, 43274337.

(9). Wi niewski JR; Zougman A; Nagaraj N; Mann M Universal sample preparation method for proteome analysis. Nat. Methods 2009, 6, 359. [PubMed: 19377485]

(10). Kessner D; Chambers M; Burke R; Agus D; Mallick P ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008, 24, 2534–2536. [PubMed: 18606607]

(11). Hoopmann MR; Finney GL; MacCoss MJ High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal. Chem 2007, 79, 5620–5632. [PubMed: 17580982]

(12). Hsieh EJ; Hoopmann MR; MacLean B; MacCoss MJ Comparison of database search strategies for high precursor mass accuracy MS/MS data. J. Proteome Res 2010, 9, 1138–1143. [PubMed: 19938873]

(13). McIlwain S; Tamura K; Kertesz-Farkas A; Grant CE; Diament B; Frewen B; Howbert JJ; Hoopmann MR; Käll L; Eng JK; MacCoss MJ; Noble WS Crux: rapid open source protein tandem mass spectrometry analysis. J. Proteome Res 2014, 13, 4488–4491. [PubMed: 25182276]

(14). Park CY; Klammer AA; Kall L; MacCoss MJ; Noble WS Rapid and accurate peptide identification from tandem mass spectra. J. Proteome Res 2008, 7, 3022–3027. [PubMed: 18505281]

(15). Kall L; Canterbury JD; Weston J; Noble WS; MacCoss MJ Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 2007, 4, 923. [PubMed: 17952086]

(16). The M; MacCoss MJ; Noble WS; Käll L Fast and accurate protein False Discovery Rates on large-scale proteomics data sets with Percolator 3.0. J. Am. Soc. Mass Spectrom 2016, 27, 1719–1727. [PubMed: 27572102]

(17). Ma Z-Q; Dasari S; Chambers MC; Litton MD; Sobecki SM; Zimmerman LJ; Halvey PJ; Schilling B; Drake PM; Gibson BW; Tabb DL IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J. Proteome Res 2009, 8, 3872–3881. [PubMed: 19522537]

(18). Cox J; Mann M MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol 2008, 26, 1367–1372. [PubMed: 19029910]

(19). Uszkoreit J; Maerkens A; Perez-Riverol Y; Meyer HE; Marcus K; Stephan C; Kohlbacher O; Eisenacher M PIA: An intuitive protein inference engine with a web-based user interface. J. Proteome Res 2015, 14, 2988–2997. [PubMed: 25938255]

(20). Fisher RA Breakthroughs in Statistics; Springer, 1992; pp 66–70.

(21). Spirin V; Shpunt A; Seebacher J; Gentzel M; Shevchenko A; Gygi S; Sunyaev S Assigning spectrum-specific p-values to protein identifications by mass spectrometry. Bioinformatics 2011, 27, 1128–1134. [PubMed: 21349864]

(22). Alves G; Yu Y-K Mass spectrometry-based protein identification with accurate statistical significance assignment. Bio-informatics 2015, 31, 699–706.

(23). Granholm V; Navarro JF; Noble WS; Kall L Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. J. Proteomics 2013, 80, 123–131. [PubMed: 23268117]

(24). Veenstra TD; Conrads TP; Issaq HJ Commentary: What to do with "one-hit wonders"? Electrophoresis 2004, 25, 1278–1279. [PubMed: 15174049]

(25). Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M A scalable approach for protein False Discovery Rate estimation in large proteomic data sets. Mol. Cell. Proteomics, 2015,14, 2394101074/ mcp.M114.046995

(26). Peng J; Elias JE; Thoreen CC; Licklider LJ; Gygi SP Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res 2003, 2, 43–50. [PubMed: 12643542]

(27). Omenn GS; Lane L; Lundberg EK; Overall CM; Deutsch EW Progress on the HUPO Draft Human Proteome: 2017 Metrics of the human proteome project. J. Proteome Res 2017, 16, 4281–4287. [PubMed: 28853897]

(28). Serang O; MacCoss MJ; Noble WS Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J. Proteome Res 2010, 9, 5346–5357. [PubMed: 20712337]

(29). Serang O; Moruz L; Hoopmann MR; Kall L Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. J. Proteome Res 2012, 11, 5586–91. [PubMed: 23148905]

(30). Audain E; Uszkoreit J; Sachsenberg T; Pfeuffer J; Liang X; Hermjakob H; Sanchez A; Eisenacher M; Reinert K; Tabb DL; Kohlbacher O; Perez-Riverol Y In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. J. Proteomics 2017, 150, 170–182. [PubMed: 27498275]

(31). Gupta N; Pevzner PA False Discovery Rates of protein identifications: a strike against the two-peptide rule. J. Proteome Res 2009, 8,4173–4181. [PubMed: 19627159]

(32). Hather G; Higdon R; Bauman A; Von Haller PD; Kolker E Estimating False Discovery Rates for peptide and protein identification using randomized databases. Proteomics 2010, 10, 2369–2376. [PubMed: 20391536]
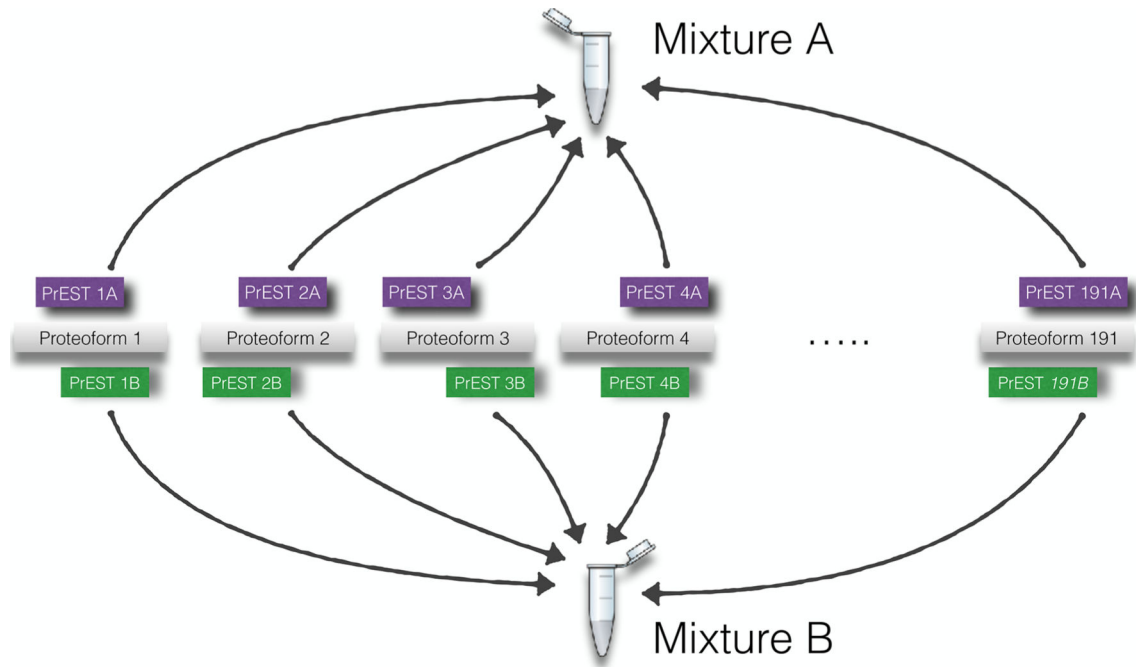
Figure 1.
The design of the two mixtures A and B. Two mixtures were generated from 191 overlapping PrEST sequences. The two mixtures, as well as a combination of the two, A+B, were mixed in a background of *E. coli*-lysate and subject for shotgun proteomics analysis.
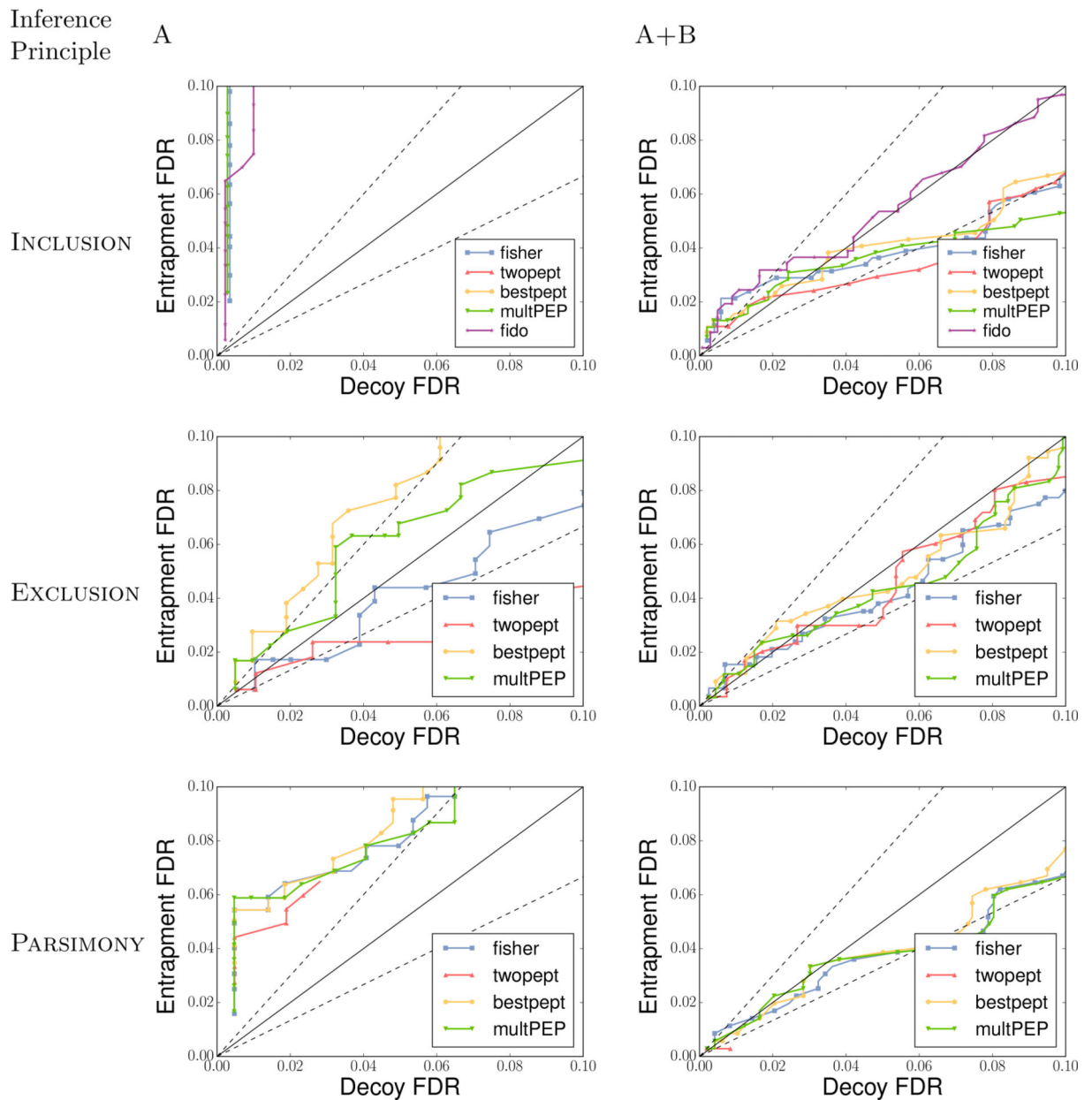
Figure 2.
Accuracy of the tested confidence estimation procedures for different inference methods. The figure plots report q values from the decoy model, the decoy FDR, against the fraction of entrapment proteins in the set of identified target proteins, with the observed entrapment FDR using a peptide-level FDR threshold of 5%. The dashed lines indicate $y = x/1.5$ and $y = 1.5x$.

Table 1.

Number of Inferred Proteins at a 5% Protein-Level Entrapment FDR from the Peptides Derived from the Triplicate Runs[a]

Inference principle	Scoring method	A	A+B
Anticipated number of PrESTs		$191 \times 1.05 = 201$	$382 \times 1.05 = 401$
Inclusion	Fisher's method	112	390
	Products of PEPs	124	395
	Best peptide	0	395
	Two peptides	0	381
	Fido	181	388
Exclusion	Fisher's method	182	345
	Products of PEPs	184	355
	Best peptide	185	355
	Two peptides	171	309
Parsimony	Fisher's method	181	365
	Products of PEPs	181	365
	Best peptide	174	365
	Two peptides	181	344

[a]Note that, since we are inferring proteins known to be present, any incorrect inferences will be added as additional proteins. Hence the maximal number of inferred proteins is 5% higher than the number of proteins in the mixtures.