Journal of proteome research

Technical Note

pubs.acs.org/jpr

Re-Fraction: A Machine Learning Approach for Deterministic Identification of Protein Homologues and Splice Variants in Large-scale MS-based Proteomics

Pengyi Yang,[†,‡,§] Sean J. Humphrey,[§,∥] Daniel J. Fazakerley,[§,∥] Matthew J. Prior,[§] Guang Yang,[§] David E. James,[*,§] and Jean Yee-Hwa Yang[*,‡]

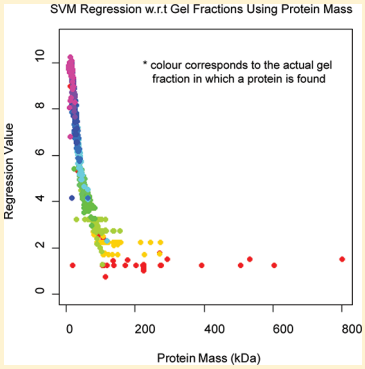
[†]School of Information Technologies, University of Sydney, NSW 2006, Australia

[‡]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

[§]Diabetes and Obesity Program, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

Supporting Information

ABSTRACT: A key step in the analysis of mass spectrometry (MS)-based proteomics data is the inference of proteins from identified peptide sequences. Here we describe Re-Fraction, a novel machine learning algorithm that enhances deterministic protein identification. Re-Fraction utilizes several protein physical properties to assign proteins to expected protein fractions that comprise large-scale MS-based proteomics data. This information is then used to appropriately assign peptides to specific proteins. This approach is sensitive, highly specific, and computationally efficient. We provide algorithms and source code for the current version of Re-Fraction, which accepts output tables from the MaxQuant environment. Nevertheless, the principles behind Re-Fraction can be applied to other protein identification pipelines where data are generated from samples fractionated at the protein level. We demonstrate the utility of this approach through reanalysis of data from a previously published study and generate lists of proteins deterministically identified by Re-Fraction that were previously only identified as members of a protein group. We find that this approach is particularly useful in resolving protein groups composed of splice variants and homologues, which are frequently expressed in a cell- or tissue-specific manner and may have important biological consequences.



KEYWORDS: *Proteomics, Machine learning, Protein Inference, Protein homologues, Splice variants, Isoforms, Mass spectrometry*

INTRODUCTION

Liquid chromatography (LC) combined with tandem mass spectrometry (LC−MS/MS) has become a popular method for large-scale, high-throughput identification and quantitation of entire proteomes from cells, tissues, organelles and organisms.[1,2] In "bottom-up" MS-based proteomics, protein mixtures are typically fractionated and/or enriched before digestion with a proteolytic enzyme such as trypsin. This produces a complex mixture of peptides that are injected into the mass spectrometer via online reversed-phase chromatography coupled to electrospray ionization.[3] The mass spectrometer resolves the eluting peptides, identifying their unique mass/charge (m/z) ratios (MS), and the most abundant MS signals are selected for isolation and fragmentation (MS/MS). The raw data output from the mass spectrometer from a complex sample contains tens of thousands of spectra including m/z information for intact peptides or "precursor ions" (MS spectra), and peptide fragment spectra (MS/MS spectra). A typical aim in the analysis of MS-based proteomics data is to identify proteins that are present within the original sample. This is typically achieved computationally through an analysis pipeline that includes peak detection,[4] database searching for peptide-spectrum matches (PSMs)[5] (Table 1), PSM validation and filtering,[6] and protein inference from identified peptides.[7]

The analysis of MS-based proteomics data is complicated by the large volumes of data generated and high sample complexity. Therefore, the development of robust and efficient computational strategies is of critical importance for improving the accuracy of protein identification[8] and quantitation.[9] One specific computational challenge is to infer which proteins are present in the sample based on the identified peptides, as the association between protein and peptide is lost once a complex sample is enzymatically digested. This problem is referred to as protein inference.[10] The key challenge faced during protein inference is the high percentage of identified peptides that can be shared among multiple proteins (shared peptides), which results in ambiguity in determining the exact identity of proteins present in the sample. This is common in higher organisms such as human and mouse due to a high degree of sequence identity between homologous proteins, protein isoforms, and/or alternative splice variants. It is estimated by

Received: January 21, 2012
Published: March 19, 2012

Table 1. Summary of Terminology

term	description
Peptide-spectrum matches	A peptide that is computationally matched by an MS/MS spectrum.
Peptide identification	A peptide that is supported by one or more PSMs.
Unique peptide	A peptide that is uniquely assigned to only one protein entry in a protein database.
Shared/degenerate peptide	A peptide that is assigned to two or more protein entries in a protein database.
Deterministic protein	A protein with at least one unique peptide assigned.
Nondeterministic protein	A protein for which all assigned peptides are shared peptides.
Minimal protein list	Proteins from protein groups which the observed peptides could be accounted for.
Protein groups	The set of all possible proteins from a minimal protein list that have one or more identified peptides in common.
Splice variant proteins	Proteins that are translated from alternative splicing of a gene.
Splice variant group (SVG)	A group of proteins that a translated from the alternative splicing of the same gene.

Meyer et al. that among 89486 human proteins recorded in the International Protein Index (v.3.75) database, >2 million of 3.8 million fully tryptic peptides are shared between two or more proteins (~53%).[11] Similarly, our analysis of the mouse proteome using International Protein Index (v.3.85) indicates that of the 59979 mouse proteins, 419633 of 763703 fully tryptic peptides (allowing ≥7 amino acids and no missed cleavages) are shared between two or more proteins (~55%).

One approach to protein inference is to combine those nondeterministic proteins into a single unit called a protein group. Popular strategies include applying Occam's razor to create a minimal protein list in which all identified peptides are accounted for;[7] utilizing a graphical representation of peptides to deduce a parsimony from which a minimal protein list is generated;[12] clustering shared peptides and deducing a minimal protein list by a greedy algorithm;[13] including a protein when it can explain a defined number of peptides that were not explained by the proteins already present in the protein group;[14] displaying all proteins but grouping them according to a "peptide-centric" view;[11,15] or allowing users to decide whether to display all proteins or only deterministic protein identifications.[16]

Limitations with protein group-based analyses arise from difficulty in determining which protein is present in the sample. This becomes a problem when combining proteomics data with other -omics data, such as transcriptomics, as only deterministic protein identifications can be confidently used. There are several proposed strategies for determining the most likely protein from each protein group. These include ranking proteins within a group with respect to the number of identified peptides or sequence coverage;[17] using a Bayesian approach;[18] using a mixture model to combine peptide identification and protein inference;[19] classifying the identified peptides using a sequence-protein-accession-gene model[20,21] and its extension using a Markov inference approach.[22] Although these methodological improvements provide a better indication of which proteins are most likely to be present in the sample, they cannot deterministically resolve the ambiguity in protein inference.

It has been suggested that protein separation techniques such as gel electrophoresis could assist the determination of the protein identity by utilizing additional information such as molecular weight and/or isoelectric point (p*I*).[10,23] Indeed, as shown by Pedersen et al., the use of gel electrophoresis to target a particular type of protein followed by manual evaluation to incorporate p*I* and mass greatly increases the power of deterministic protein identification.[24] However, making use of this information in protein inference of large-scale MS-based proteomics of complex organisms manually is impractical and may also be subjective. Therefore, an automated, efficient, and objective approach for utilizing information inherent within protein fractionation methodologies to aid protein inference would be highly advantageous for increasing the power of deterministic protein identification in large scale MS-based proteomics studies.

We propose a novel machine learning approach for deterministic protein identification that can be used to reduce the ambiguity in protein inference. This is accomplished by using a support vector machine (SVM) regression model built on proteomics data generated from samples fractionated by gel electrophoresis. We show that our algorithm accurately assigns each protein to its corresponding fraction by using a combination of four protein physical properties (i.e., mass, length, number of tryptic peptides, and p*I*). Since the fraction from which a peptide was identified is known, this information can be used to prevent the peptide from being assigned to unlikely or incorrect proteins based on their physical properties, even if all putative proteins in the protein group contain the same observed peptide sequences. We name this method Re-Fraction and a key feature of the algorithm presented is that it is computationally efficient and fully automated, being able to process proteome-scale data in the order of several minutes on typical desktop computer and requiring no manual intervention. As a result, our algorithm minimizes subjectivity and is well suited for large-scale MS-based proteomics. By analyzing previously published MS-derived proteomics data from our laboratory using Re-Fraction we demonstrate: (1) a significant improvement in the number of unique peptide assignments, and therefore the number of deterministic protein identifications; (2) the assignment of many more peptides uniquely to the proteins that have been previously deterministically identified, resulting in a higher confidence for the identified proteins; and (3) the deterministic identification of homologues and splice variant proteins that previously were only identified as part of protein groups.

■ MATERIALS AND METHODS

Benchmark Data Sets

Sample Preparation and LC–MS/MS Analysis. Raw data was obtained from Prior et al.[25] Briefly, isolated plasma membrane (PM) proteins from 3T3-L1 adipocytes were resolved by SDS-PAGE. Lanes were cut into 10 fractions and subjected to tryptic digestion.[26] For LC–MS/MS analysis, peptides were separated on a Dionex Ultimate 3000 LC system, and technical replicate analysis was performed by a LTQ-FT Ultra or Orbitrap Velos mass spectrometer. The data from each instrument are herein referred to as "LTQ-FT data set" and "Orbitrap data set". The raw mass spectrometry data analyzed in this study is available via the online repository Tranche using the hash 8W6p1KCa16W58wEF44xaNq/Xg0qGhuKfLLCYT3p8k9mCBMO/WxCGs8tlre1DQO8mtES/rvK+OdolTNiNNonfyNOhqjIAAAAAAAA5Kw==.

Data Analysis. Data were preprocessed using the MaxQuant software version 1.2.0.18 package as described previously[27] using default settings, modifications: Oxidised Methionine (M) and Acetylation (Protein N-term). Database searching was performed using the Andromeda search engine integrated into the MaxQuant environment[28] using the mouse IPI database v3.85. False discovery rate (FDR) at the peptide and protein level was determined using the target-decoy FDR.[29] The posterior error probability (PEP) for each peptide and each protein group was calculated by MaxQuant.[27] Similar to MaxQuant, we assigned PEPs for each deterministically identified protein as the product of the individual peptide PEPs. Only peptides that were uniquely assigned to a protein were used in the calculation of protein PEP. Peptide identifications, deterministic protein identifications, and protein group identifications were sorted by their corresponding PEP controlling for a 1% FDR, at the peptide, protein group and deterministic protein levels.

Proteins that were deterministically identified from MaxQuant results are referred to as "original deterministic protein" identifications. Proteins deterministically identified only as a result of using Re-Fraction are referred to as "additional deterministic protein" identifications.

Data Set Construction and Model Learning

We propose a two-stage framework for correcting peptide-to-protein assignment using fraction information, for mass spectrometry data in which a protein-level fractionation (e.g., SDS-PAGE) has been performed and the fractions analyzed separately. The first stage involves constructing a modeling data set and building a classification model to establish the protein−fraction relationship. The second stage is to apply the model for the correction of shared peptide assignments.

Constructing Modeling Data set

There are three steps for constructing the modeling data set to build the classification model. First, peptide identifications from MaxQuant are filtered to select training samples. Second, we determine the fraction relationship for each protein by assigning each to a most appropriate fraction. Third, we extract protein properties and their MS statistics so as to create learning features. Figure 1 illustrates the filtering and labeling steps involved. The training data construction and the model learning are done uniquely for any given data set (in our case, LTQ-FT and Oribtrap data sets, respectively). Re-Fraction is therefore flexible and highly specific and should be widely applicable for any large proteomics data set utilizing protein-level fractionation.

Step 1: Peptide Filtering. We used the following hierarchical classification criteria to filter peptide identifications for model training:

1. Only peptides uniquely assigned to one protein are selected. Figure 1 shows 20 peptides each uniquely assigned to only one protein.
2. Within those uniquely assigned peptides, only those that have PSMs supporting their identification in adjacent fractions are selected. In Figure 1, peptide "P17" failed to pass this criterion because the supporting PSMs are from disjoint fractions (i.e., "F1", "F2" and "F4").
3. For those that passed criteria 1 and 2, only those that have PSMs supporting their identification in a maximum of 3 fractions are selected. In Figure 1, peptide "P16" and "P18" failed to pass this criterion since they have PSMs reported from more than 3 fractions.

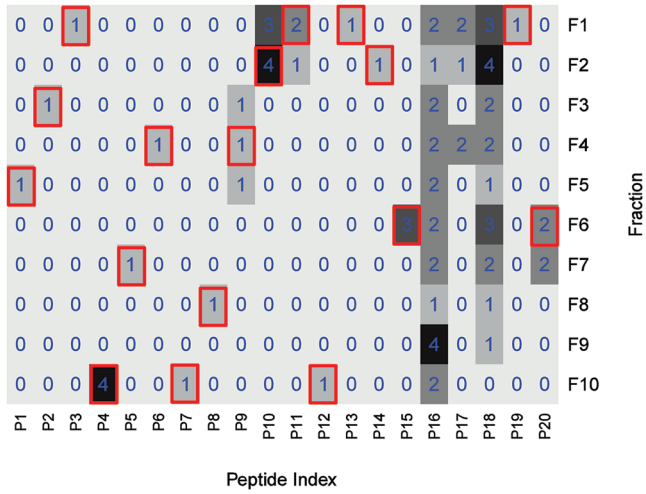


Figure 1. Illustration on construction of the training data set. The *x*-axis is an index of 20 peptides each uniquely assigned to one protein. The *y*-axis is the 10 gel fractions. The color depth and the number in each locus indicate the number of spectra matched to a peptide. For example, peptide "P10" has PSMs identified from fraction "F1" (3 spectral counts) and "F2" (4 spectral counts).

Step 2: Protein Fraction Labeling. For those peptides that passed the filtering criteria, the protein-fraction relationships are determined as follows:

1. For the peptides that have PSMs identified in only one fraction, the corresponding protein is assigned to this fraction. In Figure 1, the protein corresponding to peptide "P1" is assigned to fraction "F5".
2. For the peptides that have PSMs identified in two fractions, the corresponding protein is assigned to the fraction with the largest number of spectrum counts. In Figure 1, the protein corresponding to peptide "P10" is assigned to fraction "F2". If the number of spectrum counts from the two fractions is the same, the fraction that corresponds to a larger molecular weight is assigned. In Figure 1, the protein corresponding to peptide "P20" is labeled to fraction "F6".
3. For the peptides that have PSMs identified in three fractions, the protein is assigned to the fraction with the largest number of spectrum counts as in 2. If the number of spectrum counts from the three fractions is the same, the protein is assigned to the center fraction. For example, the protein corresponding to peptide "P9" in Figure 1 is assigned to "F4".

Step 3: Extracting Protein Features. For proteins for which fraction relationships could be determined, we extracted 4 features for modeling the protein-fraction relationship. These features are: protein mass (kDa), protein length (number of amino acids), number of the theoretical tryptic peptides, theoretical p*I*. For LTQ-FT data set, proteins are visually resolved into distinct fractions on the basis of each of the 4 features (Figure 2). Similar separation patterns were observed using the Orbitrap data set (data not shown). The separation of the 4 extracted features were evaluated in an assessment of the model, by using protein mass alone and by combining two, three, and all four features.

dx.doi.org/10.1021/pr300072j | *J. Proteome Res.* 2012, 11, 3035−3045

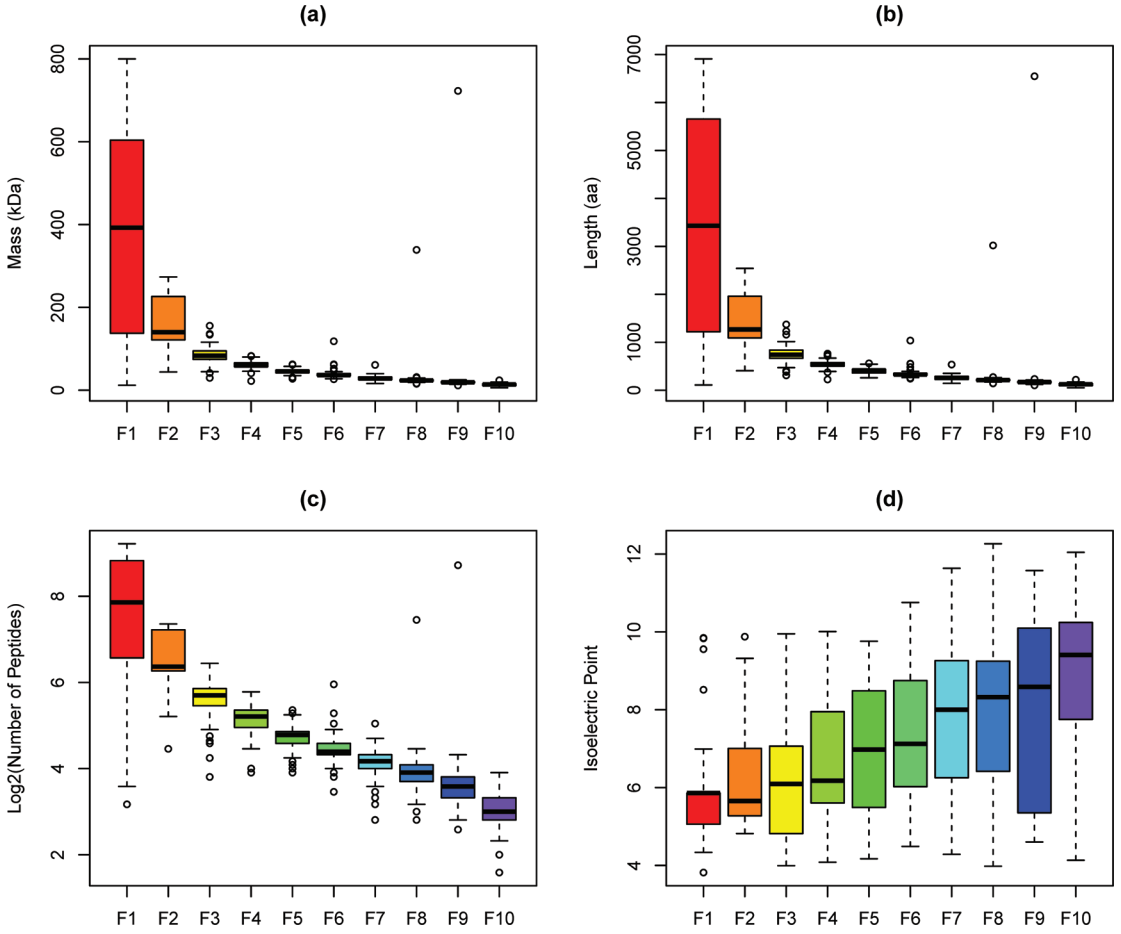


Figure 2. Separation of proteins from LTQ-FT data set. Proteins were separated with respect to (a) mass (kDa); (b) length (number of amino acids); (c) number of theoretical tryptic peptides (log2); (d) theoretical p*I*. The *x*-axis is the index of gel fractions. Each color denotes a gel fraction.

Modeling the Protein–Fraction Relationship

We used the LibSVM Support Vector Machine (SVM) library[30] for constructing a model for protein to fraction classification. Specifically, we used regression mode, since the fractions are artificially imposed on a continuous gel-electrophoresis separation. The default kernel of radial basis function was used. First, we assessed the performance of the model using a stratified 10-fold cross validation. We then generated a final model using the entire data set for correcting shared peptide assignments.

Model Assessment. We evaluated the model using a stratified 10-fold cross validation on the modeling data where equal proportion of proteins from each fraction was divided into 10 segments. Nine segments were used for model training, while the last segment was reserved for model assessment. This was repeated until each of the 10 segments had been used for model assessment. To classify the proteins to their corresponding fraction, the regression value for each testing instance was rounded to an integer. This allowed us to compare whether the SVM predicted fractions that corresponded to the actual fraction. In evaluating different combinations of features, we report the average accuracy (Acc), sensitivity (Se) and specificity (Sp) across all fractions computed from the 10-fold cross validation procedure. For the combination of all four features, we divide and report the average accuracy, sensitivity and specificity with respect to each fraction.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{Se} = \frac{\mathrm{TP}P}{\mathrm{TP}P + \mathrm{FN}N}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative predictions, respectively.

Final Model. The stratified 10-fold cross validation provides a measure of performance indicating how well the SVM correctly assigns proteins to their corresponding gel fractions. However, it does not determine a final model, which is needed for correcting shared peptide assignments. To derive the final model we therefore used the entire modeling data set.

After obtaining the final model, where the same peptide was assigned to multiple proteins, we classified each protein to a fraction and checked for PSMs in the fraction that contributed to the peptide identification. If the spectrum count was zero, the protein was removed from the assignment. Otherwise, the peptide-to-protein assignment was retained. The results obtained in the original analysis (referred to as "original") are compared to those generated using Re-Fraction.

dx.doi.org/10.1021/pr300072j | *J. Proteome Res.* 2012, 11, 3035–3045

Evaluation

For peptide results comparison, we compared the number of unique peptide assignments from the original result generated by MaxQuant with that from Re-Fraction. For the protein results, we applied four evaluation methods. First, we compared the total number of deterministic protein identifications from the original result with that from using Re-Fraction. Second, we examined the number of additional deterministic protein identifications in each fraction by using Re-Fraction. Those additional deterministic protein identifications are evaluated bioinformatically using gene ontology (GO) enrichment analysis.[31] Specifically, we applied a hyper-geometric test with respect to the GO terms "membrane" (GO:0016020) and "nucleus" (GO:0005634). Since the plasma membrane (PM) proteins were enriched before MS analysis, we expect to identify significantly more proteins from the PM as opposed to an organelle not associated with the PM (nucleus).

The third evaluation is to obtain all proteins from the minimal protein list for LTQ-FT and Orbitrap results, respectively, and overlap the additional deterministic protein identifications from the LTQ-FT data set with the proteins from the minimal protein list of Orbitrap and *vice versa*. We considered the additional deterministically identified proteins reasonable if they could be found in the minimal protein list of a technical replicate generated by using a different instrument analyzing the same sample.

The fourth evaluation is that for each deterministic protein identification in the original result generated by MaxQuant, we calculated how many additional unique peptides were assigned to them after using Re-Fraction.

Finally, we compiled a high confidence list of additional deterministic proteins from each of the LTQ-FT and Orbitrap data sets using Re-Fraction. These proteins were selected on the basis that they were not found in gel fractions adjacent to the other proteins within the same protein group. We then manually verified and classified those identifications according to whether they are resolved from completely different proteins or alternative splice variant proteins using the UniProt database.[32]

SDS-PAGE, Immunoblotting, and Protein Structure Analysis

All samples were subjected to SDS-PAGE analysis on 10% resolving gels. Equal amounts of protein were loaded for each sample in a single experiment with 10 μg/lane. Separated proteins were electrophoretically transferred to PVDF membrane, blocked with 5% nonfat skim milk in 0.1% Tween 20 in TBS (TBST), and incubated with primary antibody in 5% BSA in TBST overnight at 4 °C. After incubation, membranes were washed three times in TBST and incubated with HRP-labeled secondary antibodies in 5% nonfat skim milk in TBST for 1 h. Proteins were visualized using Supersignal West Pico chemiluminescent substrate and imaged with X-ray film (Fuji). RagA antibody (D8B5) was purchased from Cell Signaling Technology.

Protein structures of RagA and RagB were obtained from The Protein Model Portal database.[33] SWISSMODEL[34] was selected as the model provider and the template used to create the 3-D structure was "3r7wC".

In-silico Simulation

Computational Analysis of Splice Variant Proteins in the Mouse Proteome. We annotated and selected splice variant proteins from the mouse proteome database (IPI v.3.85) using the UniProt definition of alternative splice variants/isoforms. We then grouped these splice variant proteins into splice variant groups (SVGs), where each group contains alternative splice variant proteins generated from the same gene. We calculated how many SVGs are distinguishable (i.e., where two splice variants within an SVG can be distinguished from one another) at a range of mass cutoffs, including those that are achievable using the current gel fractionation.

Computational Analysis of Shared Peptides in the Mouse Proteome. We tryptically digested the mouse proteome *in-silico* using International Protein Index (v.3.85) database. Fully tryptic peptides with ≥7 amino acids and no missed cleavages were accepted. We calculated the percentage of shared peptides that could be uniquely resolved given a 2D protein separation with respect to protein mass and protein p*I* at a given resolution.

Taking mass calculation as an example, for a peptide that is assigned to multiple proteins, we obtain the median mass of those proteins. Then, given a resolution, *d*, we calculate a range of median mass plus/minus *d* and exclude those proteins whose masses fall outside the range. If only one peptide-to-protein assignment remains after exclusion, then at the resolution, *d*, this peptide can be resolved to be a unique assignment. This is done similarly for different p*I* fractionation resolution and the results of the two are combined to indicate the power of 2D protein separation on unique peptide assignment.

RESULTS

Model Performance

The use of an SVM to classify a given protein to its corresponding gel fraction is central to our rapid and automated approach to remove potentially false peptide-to-protein assignments. If the SVM model used is accurate, we can accurately reduce the ambiguity of multiple peptide assignments and increase unique peptide assignments. Therefore, Re-Fraction can be used to increase the number of deterministic protein identifications in large-scale proteomics data sets.

We evaluated the performance of the SVM using one or more protein features for protein−fraction classification. Table 2 shows the performance of the SVM model in terms of average accuracy, sensitivity, and specificity across all fractions from the stratified 10-fold cross validation using one or more features. The use of several protein features resulted in improved

Table 2. SVM Model Evaluation using One or More Features in Predicting Protein−Fraction Relationships in Terms of Accuracy (Acc), Sensitivity (Se), and Specificity (Sp) from 10-Fold Cross Validation

features	Acc	Se	Sp
LTQ-FT data set			
Mass	0.949	0.73	0.971
Mass and numTryptic	0.95	0.737	0.972
Mass and numTryptic and Length	0.951	0.74	0.972
Mass and numTryptic and Length and p*I*	0.968	0.814	0.982
Orbitrap data set			
Mass	0.942	0.696	0.968
Mass and numTryptic	0.944	0.699	0.968
Mass and numTryptic and Length	0.948	0.726	0.971
Mass and numTryptic and Length and p*I*	0.963	0.79	0.979

dx.doi.org/10.1021/pr300072j | J. Proteome Res. 2012, 11, 3035−3045

Table 3. SVM Model Evaluation using All Four Features in Predicting Protein−Fraction Relationships[a]

fraction	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
					LTQ-FT data set					
Acc	0.981	0.988	0.986	0.967	0.955	0.957	0.946	0.952	0.964	0.981
Se	0.824	0.936	0.944	0.937	0.863	0.801	0.753	0.628	0.722	0.737
Sp	0.999	0.994	0.992	0.975	0.968	0.978	0.968	0.975	0.975	0.996
					Orbitrap data set					
Acc	0.979	0.972	0.971	0.953	0.959	0.952	0.942	0.95	0.966	0.983
Se	0.805	0.862	0.916	0.917	0.849	0.678	0.735	0.717	0.664	0.755
Sp	0.998	0.987	0.98	0.962	0.972	0.979	0.968	0.968	0.982	0.996

[a]Performance is reported for each fraction in terms of accuracy (Acc), sensitivity (Se), and specificity (Sp) from 10-fold cross validation.

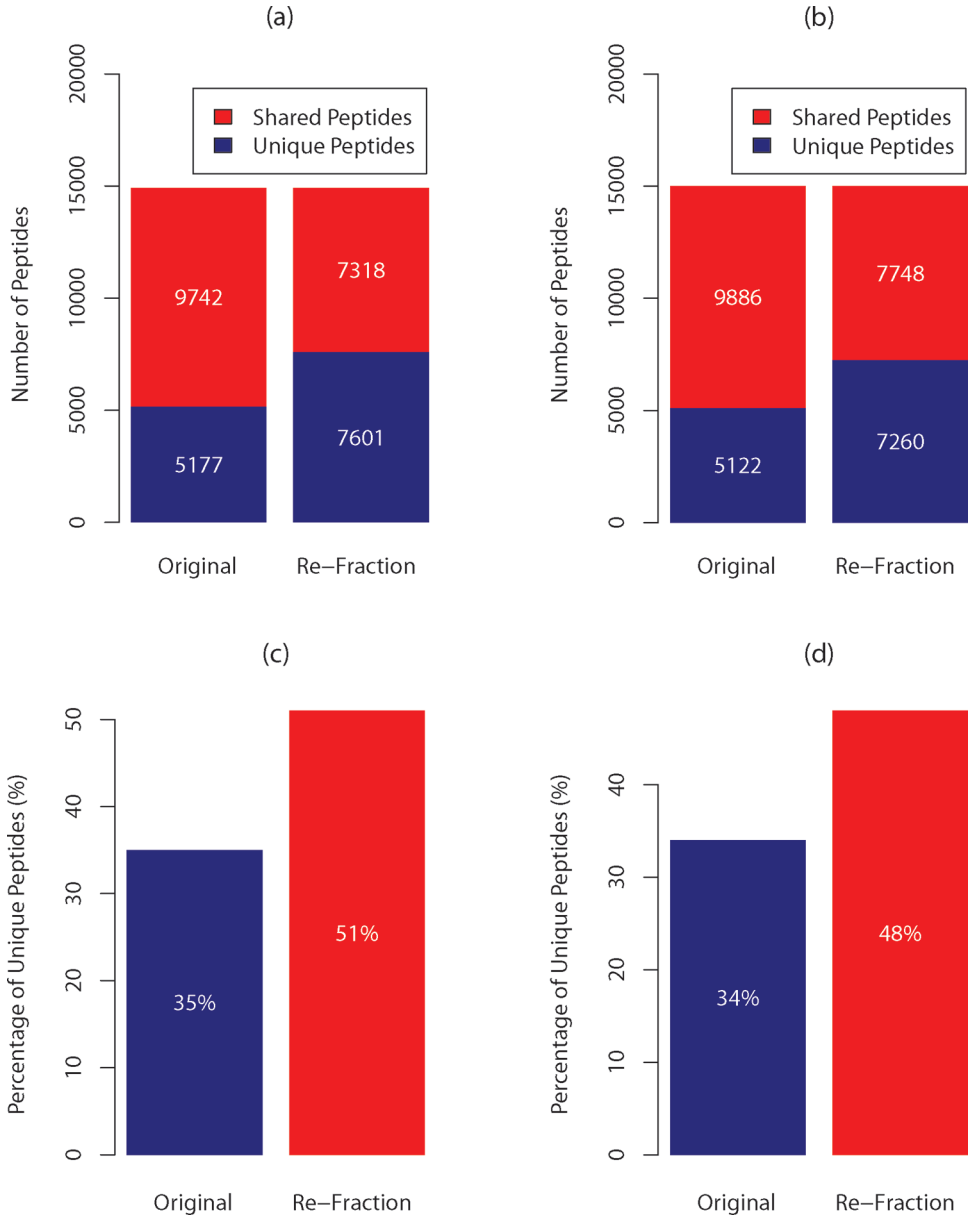


Figure 3. Peptide assignment for LTQ-FT data set and Orbitrap data set. The peptide assignments were categorized into unique peptides (blue) and shared peptides (red) for (a) LTQ-FT and (b) Orbitrap data sets. Percentage of unique peptide assignments with respect to all identified peptides in (c) LTQ-FT and (d) Orbitrap data sets.

predictive performance of the SVM than using mass alone for both LTQ-FT and Orbitrap data sets (Table 2).

Table 3 shows the performance of the SVM using all four features. The performance in terms of accuracy and specificity was high across most fractions for both LTQ-FT and Orbitrap data sets. The sensitivity appeared to tail off at both ends of the gel while specificity remained relatively stable across all fractions. This drop in sensitivity may be due to the lower

dx.doi.org/10.1021/pr300072j | J. Proteome Res. 2012, 11, 3035−3045

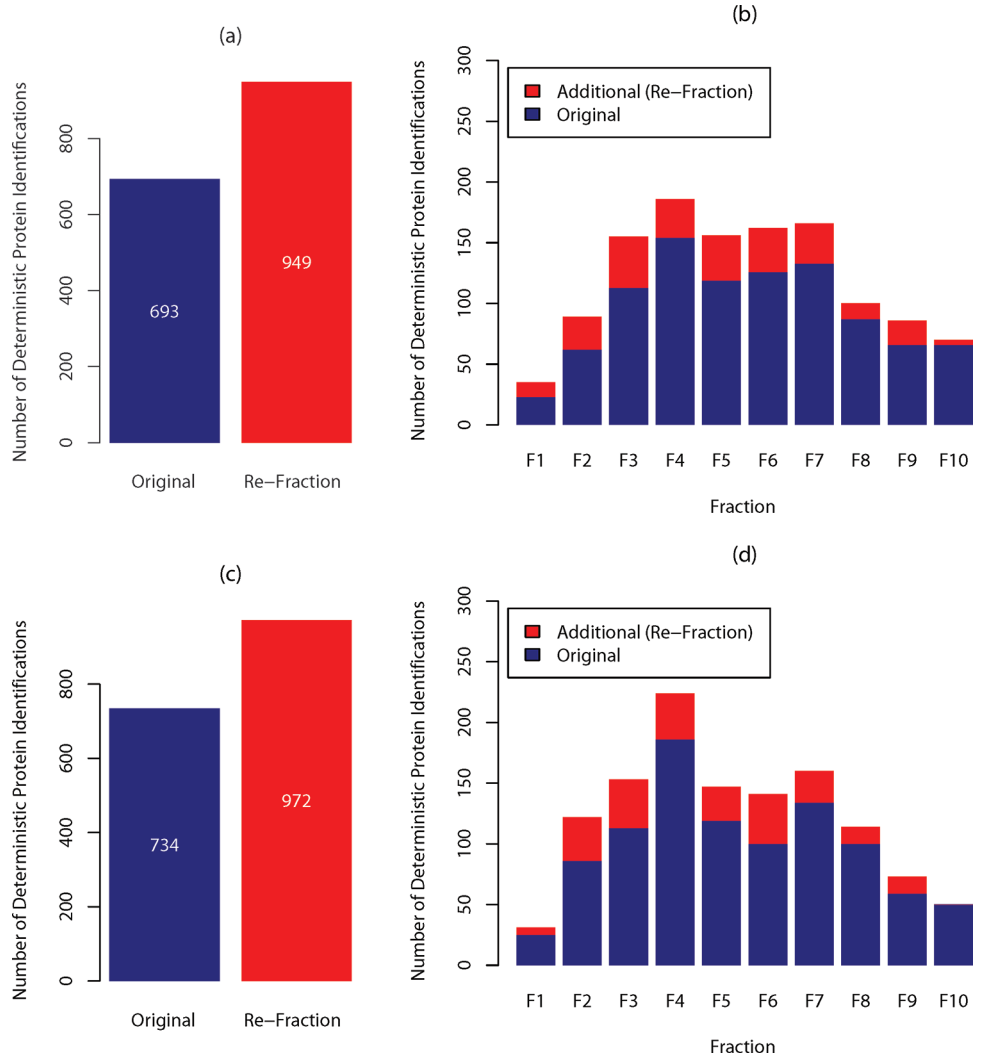


Figure 4. Deterministic protein identifications. Total number of deterministic protein identifications for (a) LTQ-FT and (c) Orbitrap data sets, respectively. The deterministic protein identifications are further divided with respect to the gel fractions for (b) LTQ-FT data set and (d) Orbitrap data set. Blue segments correspond to the deterministic identifications from the original analysis, while red segments correspond to the additional deterministically identified proteins resulting from applying Re-Fraction.

resolving power of the polyacrylamide gel for these molecular weight ranges, since the gel used in this data set was a fixed acrylamide concentration, and may also be affected by the relatively fewer number of proteins identified in these fractions.

Peptide Assignment

By applying Re-Fraction to process our MS-based proteomic data, we have substantially increased the number of unique peptides assigned. A total of 14,919 and 15,008 peptides were identified by controlling for 1% peptide FDR for LTQ-FT and Orbitrap data sets, respectively. For the LTQ-FT data set, the use of Re-Fraction increased the number of unique peptides from 5177 to 7601 (Figure 3a), a ~47% increase. For the Orbitrap data set, the number of unique peptides increased from 5122 to 7260 by using Re-Fraction (Figure 3b), a ~42% improvement. The increase achieved was 16 and 14% with respect to all identified peptides in LTQ-FT and Orbitrap data sets, respectively (Figure 3c,d). These results demonstrate that a highly consistent improvement in unique peptide assignment can be achieved using Re-Fraction. Note that using Re-Fraction ~50% of the identified peptides could still not be uniquely assigned to a protein with current gel resolution, implying that there are still a large number of proteins that cannot deterministically identified.

Deterministic Protein Identification

At the protein level, the use of Re-Fraction on the LTQ-FT and Orbitrap data sets substantially increased the number of deterministic protein identifications as a result of increased unique peptide assignments. For the LTQ-FT data set, 949 proteins were deterministically identified compared to the 693 proteins from the original approach at 1% FDR, an increase of 37% (Figure 4a). This represents 256 additional deterministic identifications by using Re-Fraction. For these 256 additional deterministic protein identifications, a gene ontology (GO) term enrichment analysis yielded a p-value of 2×10^{-3} with respect to the GO term "membrane" (GO:0016020) while an enrichment p-value of 0.97 was obtained with respect to the GO term "nucleus" (GO:0005634).

Similarly, for the Orbitrap data set, Re-Fraction increases the number of deterministic protein identifications from 734 to 972 with an FDR of 1%, a 32% increase (Figure 4c). Additional identification (243) were highly enriched with respect to the GO term "membrane" (p-value = 3×10^{-4}) and again no

dx.doi.org/10.1021/pr300072j | J. Proteome Res. 2012, 11, 3035–3045

enrichment was found with respect to the term "nucleus" (p-value = 0.76). Given the nature of this proteomics data set (PM enrichment) these findings suggest that the additional deterministic identifications resulting from the use of Re-Fraction are biologically relevant.

By looking at the number of deterministic protein identifications within each fraction, it became evident that there was reduced resolution in the upper and lower fractions (Figure 4b,d). For those fractions containing more protein identifications, Re-Fraction generated a greater number of additional deterministic protein identifications. This finding suggests that the use of Re-Fraction is likely to be particularly beneficial when larger amounts of data are available.

Those additional deterministic protein identifications from Re-Fraction were validated by overlapping them with the minimal list proteins obtained from the technical replicate. Almost all new deterministically identified proteins from the Orbitrap data set could be found in the minimal protein list from the LTQ-FT data set and vice versa by filtering minimal protein list with 1% protein groups FDR (Figure 5).

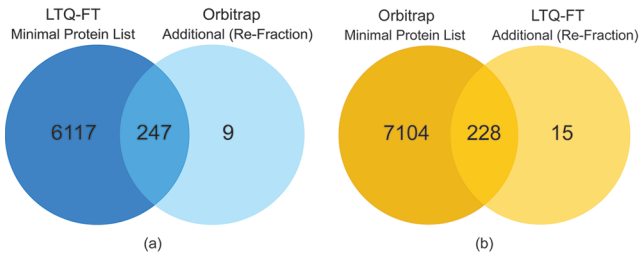


Figure 5. Validation of new deterministic protein identifications from Re-Fraction. The proteins of the minimal protein list from the LTQ-FT data set were obtained and overlapped with the new deterministic protein identifications using Re-Fraction with the (a) Orbitrap data set and (b) vice versa.

After filtering at a 1% protein level FDR there were 693 (LTQ-FT) and 734 (Orbitrap) deterministically identified proteins from the original MaxQuant results. Of these, 23% received one or more additional unique peptide assignments from using Re-Fraction in both data sets (Figure 6). For those

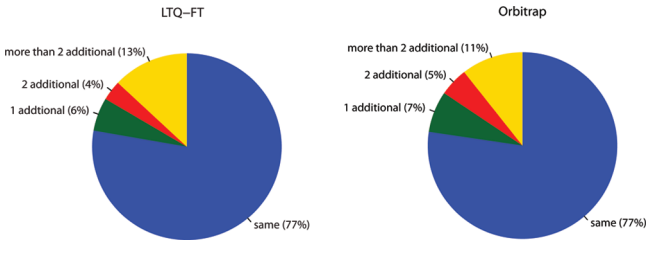


Figure 6. Additional unique peptide identifications assigned to the original deterministically identified proteins using Re-Fraction.

proteins that received additional unique peptide assignments there was an increased confidence of protein identification and these proteins may also have improved quantitation due to a larger number of available peptides.

Lastly, we compiled a high confidence list of additional deterministic proteins using Re-Fraction on each of the LTQ-FT and Orbitrap data sets (Supplementary Table 1, Supporting Information). This generated 41 and 42 additional deterministic proteins from LTQ-FT and Orbitrap data sets, respectively, that met the filtering criteria (see Methods and Materials). For the LTQ-FT data set, three were resolved from different proteins (7%), 17 were from isoforms verified at the protein (27%) or transcript (15%) level, while the remaining 21 (51%) were derived from fragment sequences predicted from DNA open reading frames (ORFs). For the Orbitrap data set, four were resolved from different proteins (9%), 17 were from isoforms verified at the protein (24%) or transcript (17%) level, and the remaining 21 (50%) from fragment sequences predicted from DNA ORFs.

Validation and Functional Implications

In our additional protein identification lists (Supplementary Table 1, Supporting Information), one example in distinguishing highly related proteins by using Re-Fraction was the identification of RagA but not RagB in adipocytes. Rags are Ras-related GTP-binding proteins that play a key role in the activation of mTOR.[35] The family is comprised of four homologous Rags A–D.[36] RagA and RagB are highly homologous proteins. They differ by seven conservative amino acid substitutions, and 33 additional residues at the N terminus of RagB[37] that encodes a small beta-sheet and two long random coils that extend an α helix chain away from the main structure. These features can be visualized by protein structure analysis (Figure 7). Based on these differences RagB is

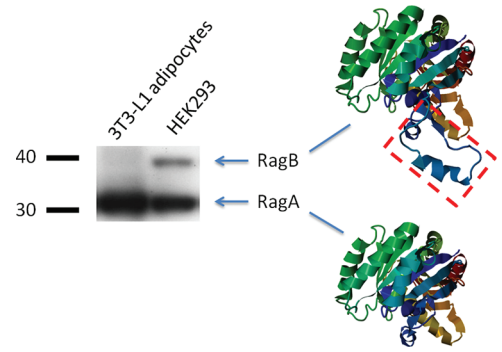


Figure 7. 3T3-L1 adipocytes express RagA but not RagB, as shown by immunoblotting. 3T3- L1 adipocytes and HEK cells were lysed and lysates were immunoblotted with an antibody that recognizes both RagA and RagB. Protein structure analysis shows an additional α helix chain (marked by a red rectangle).

predicted to be 7 kDa larger than RagA. To validate the prediction from Re-Fraction that we had identified only RagA in adipocytes, we performed immunoblotting on either 3T3-L1 adipocyte or HEK cell lysates using a Rag antibody that recognizes both homologues. As shown in Figure 7, HEK cells express both RagA and RagB while only RagA was evident in adipocytes. This provides a striking example of the utility of this new method and shows how it can used to dissect important novel functional information.

As an example of the ability of Re-Fraction to distinguish between alternative splice variant proteins, the protein "Naca Nascent polypeptide-associated complex subunit alpha (or "Naca-201" as denoted in Ensembl database)" was found to be present in the adipocytes rather than its alternative splice variant "Naca Nascent polypeptide-associated complex subunit alpha, muscle-specific form (or "Naca-202 as in Ensembl database")". Figure 8 shows a protein coding view of the Naca gene. Specifically, there is a large deletion of amino acid sequence for Naca-201 whereas Naca-202 is much longer in

dx.doi.org/10.1021/pr300072j | J. Proteome Res. 2012, 11, 3035–3045

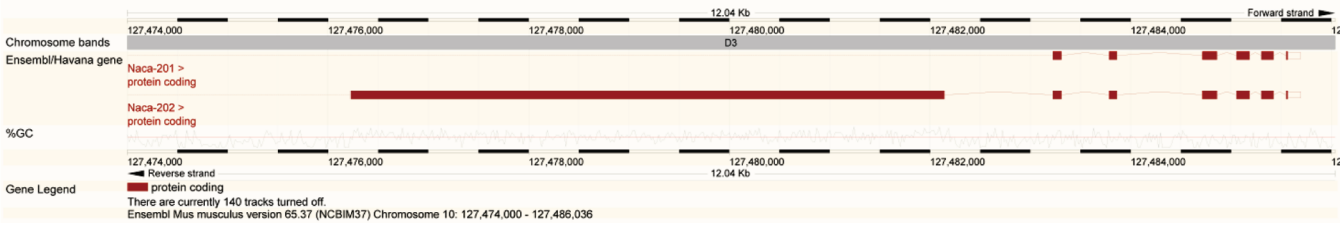


Figure 8. Protein code view of Naca gene. Two alternative splice variant proteins (denoted as "Naca-201" and "Naca-202" in Ensembl database) are coded by Naca gene. Using Re-Fraction, only "Naca-201" is found to be present in the sample, whereas "Naca-202", which is a muscle-specific form, is excluded.

amino acid sequence. As a result these splice variants have very different masses (23 kDa versus 220 kDa). According to the tissue specificity meta-data in the UniProt database, Naca-201 is ubiquitously expressed while Naca-202 is only expressed in muscle tissue. Since the proteomics data set analyzed in this manuscript was generated from the 3T3-L1 adipocyte cell line, this is an expected result. This example further demonstrates the biological utility of Re-Fraction in distinguishing between specific alternative splice variant proteins.

In-silico Simulation Using Mouse Proteome

Analysis of Splice Variant Proteins. We selected and annotated alternative splice variant proteins from the mouse proteome database (IPI v.3.85) using the UniProt definition of alternative splice variants/isoforms. Out of 54050 proteins defined in the UniProt mouse proteome, a total of 12257 proteins were annotated as having one or more alternative splice variant counterparts. These fell into 4556 splice variant groups (SVGs). Within these 4556 SVGs, 2852 (63%) were comprised of two alternative splice variant proteins, 1013 (22%) were comprised of three alternative splice variant proteins, and only 691 (15%) were comprised of 4 or more alternative splice variant proteins.

Figure 9 shows the number of SVGs that have a mass difference larger than a given mass cutoff. The figure is colored in light blue from mass cutoff of 50 to 10 (kDa), denoting the mass resolutions that may be achieved by current gel fractionation procedures. The mass cutoff from 10 to 0 (kDa) is colored in pink, denoting the mass resolutions that may be difficult or impossible to achieve by gel electrophoresis. Specifically, 1895 (42%) SVGs contain variants with a mass difference larger than a mass of 15 kDa. In the analysis of LTQ-FT and Orbitrap data sets, we identified a total of 6364 and 7332 proteins, respectively, from which there are 1181 and 1420 splice variants that fall into 481 and 586 SVGs. Using Re-Fraction, 256 additional proteins for the LTQ-FT data set and 238 proteins for the Orbitrap data set were deterministically identified. The percentages with respect to SVGs are 53% and 40% for LTQ-FT and Orbitrap data sets, respectively, indicating a reasonable agreement between theoretical performance of the method and the real-world performance.

Analysis of Shared Peptides. We assessed the theoretical performance of Re-Fraction in unique peptide assignment by performing an *in-silico* 2D protein fractionation and resolution evaluation on resolving shared peptides. Figure 10 demonstrates the theoretical bounds for tryptic peptides shared among multiple proteins, and the improvement in unique peptide assignment achievable given a prior protein separation with a given mass and p*I* resolving power.

For example, given a protein separation mass resolution of 60 kDa and p*I* resolution of 1 pH, the percentage of theoretical

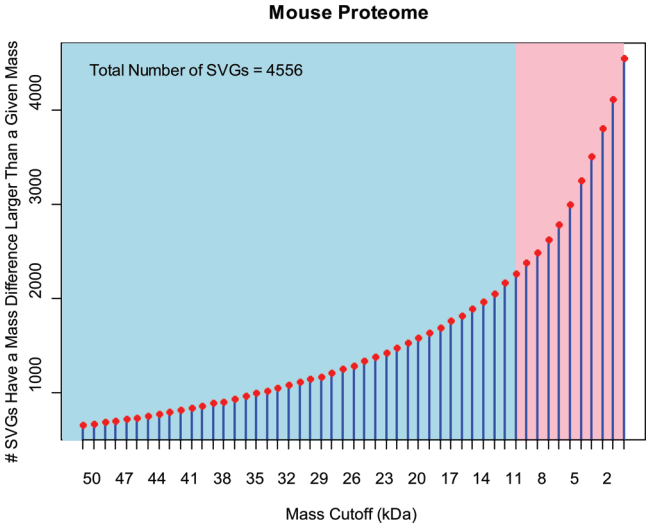


Figure 9. Analysis of alternative splice variant proteins in the mouse proteome. There are 12257 proteins that fall into 4556 splice variant groups (SVGs) where each group contains multiple splice variant proteins coded by the same gene. The *y*-axis shows how many SVGs have a mass difference larger than a given mass cutoff and *x*-axis is the given mass cutoff.

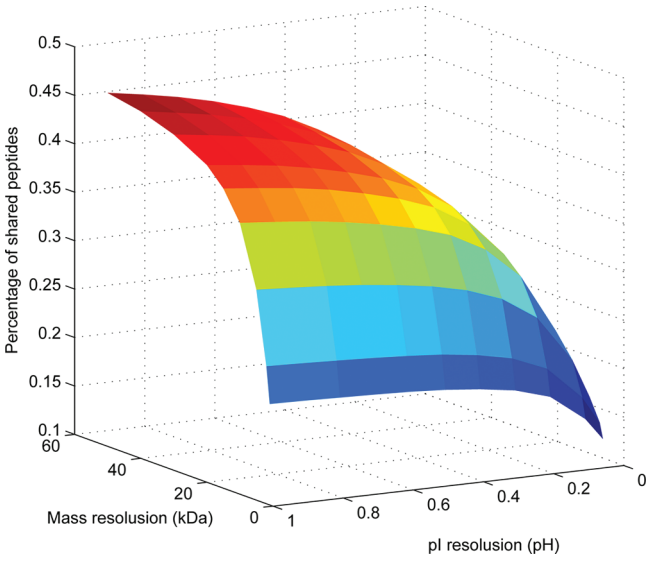


Figure 10. The mouse proteome was tryptically digested *in-silico* and the percentage of theoretical tryptic peptides shared by multiple proteins were calculated with respect to a given resolution provided by a combination of protein mass separation and protein p*I* separation.

dx.doi.org/10.1021/pr300072j | *J. Proteome Res.* 2012, 11, 3035−3045

tryptic peptides shared by multiple proteins is ~45%. By using a mass resolution of 20 kDa and p*I* resolution of 0.4 pH, the percentage reduces to around 20%. This demonstrates that by combining protein fractionation methods such as mass and charge, the percentage of shared peptides decrease exponentially. Since Re-Fraction relies on the prior protein separation for peptide assignment, intensive protein separation such as the combination of mass and charge could result in many more peptides to be uniquely assigned to a single protein, and therefore, significantly more deterministic proteins identifications.

CONCLUSION

In this study, we describe a novel machine learning approach for deterministic protein identification, called Re-Fraction. Re-Fraction reduces the ambiguity in protein inference by automatically determining the correctness of each peptide-to-protein assignment using additional information from gel electrophoresis. This is achieved by learning the protein−fraction relationship using a few protein physical properties, and subsequently correcting for peptides that have been assigned to unlikely or incorrect proteins based on which specific fraction each peptide was identified. The result is greatly improved unique peptide assignments and deterministic protein identifications accomplished in an efficient and automated way.

The proposed machine learning approach is general and can be applied to any large-scale proteomic studies where proteins are fractionated prior to LC−MS/MS analysis. Furthermore, the principle behind Re-Fraction could be adapted to other protein identification pipelines such as ProteinProphet,[7] and is not exclusive to MaxQuant. Our data indicate that many splice variant proteins and homolgos with large mass differences can be deterministically identified from their original protein groups. We have provided examples of the splice variants of "Naca" gene and homologues of protein RagA and RagB. The splice variant proteins and homologues are often biologically important as they are uniquely expressed in certain cell types and/or in response to certain treatments. Therefore, Re-Fraction not only improves deterministic protein identification but also distinguishes biologically important proteins from which new hypotheses and followup validations could be derived.

ASSOCIATED CONTENT

Supporting Information

Supplementary table. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jean.yang@sydney.edu.au; d.james@garvan.org.au. Phone: +61 (2)9351 3012. Fax: +61 (2)9351 4534.

Author Contributions

[‖]These authors contributed equally to this work.

Notes

Conflict of Interest

The authors declare that they have no competing interests.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

P.Y. is supported by the NICTA International Postgraduate Award (NIPA) and the NICTA Research Project Award (NRPA). D.F. is a Sir Henry Wellcome Post-Doctoral Fellow of the Wellcome Trust. D.E.J. is an NHMRC Senior Principal Research Fellow. This work is funded by an ARC Discovery Grant DP0984267 (J.Y.-H.Y.) and an NHMRC program grant (D.E.J.). We thank Dr. Jacqueline Stoeckli for helpful discussions.

REFERENCES

(1) Adachi, J.; Kumar, C.; Zhang, Y.; Mann, M. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol. Cell. Proteomics* **2007**, *6*, 1257−1273.

(2) Kruger, M.; Moser, M.; Ussar, S.; Thievessen, I.; Luber, C.; Forner, F.; Schmidt, S.; Zanivan, S.; Fassler, R.; Mann, M. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **2008**, *134*, 353−364.

(3) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246*, 64−71.

(4) Wang, P.; Yang, P.; Arthur, J.; Yang, J. A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data. *Bioinformatics* **2010**, *26*, 2242−2249.

(5) Eng, J.; McCormack, A.; Yates, J. III 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*, 976−989.

(6) Keller, A.; Nesvizhskii, A.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(7) Nesvizhskii, A.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646−4658.

(8) Nesvizhskii, A.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787−797.

(9) Mueller, L.; Brusniak, M.; Mani, D.; Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **2008**, *7*, 51−61.

(10) Nesvizhskii, A.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4*, 1419−1440.

(11) Meyer-Arendt, K.; Old, W.; Houel, S.; Renganathan, K.; Eichelberger, B.; Resing, K.; Ahn, N. IsoformResolver: A peptide-centric algorithm for protein inference. *J. Proteome Res.* **2010**, *10*, 3060−3075.

(12) Zhang, B.; Chambers, M.; Tabb, D. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6*, 3549−3557.

(13) Koskinen, V.; Emery, P.; Creasy, D.; Cottrell, J. Hierarchical clustering of shotgun proteomics data. *Mol. Cell. Proteomics* **2011**, *10*, M110.003822.

(14) Ma, Z.; Dasari, S.; Chambers, M.; Litton, M.; Sobecki, S.; Zimmerman, L.; Halvey, P.; Schilling, B.; Drake, P.; Gibson, B.; Tadd, D. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2009**, *8*, 3872−3881.

(15) Resing, K.; Meyer-Arendt, K.; Mendoza, A.; Aveline-Wolf, L.; Jonscher, K.; Pierce, K.; William, M.; Cheung, H.; Russell, S.; Wattawa, J.; Goehle, G.; Knight, R.; Ahn, N. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76*, 3556−3568.

(16) Tabb, D.; McDonald, W.; Yates, J. III DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1*, 21−26.

(17) Allet, N.; et al. In vitro and in silico processes to identify differentially expressed proteins. *Proteomics* **2004**, *4*, 2333−2351.

(18) Li, Y.; Arnold, R.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A Bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* **2009**, *16*, 1183−1193.
(19) Li, Q.; MacCoss, M.; Stephens, M. A nested mixture model for protein identification using mass spectrometry. *Ann. Appl. Stat.* **2010**, *4*, 962−987.
(20) Qeli, E.; Ahrens, C. PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat. Biotechnol.* **2010**, *28*, 647−650.
(21) Grobei, M.; Qeli, E.; Brunner, E.; Rehrauer, H.; Zhang, R.; Roschitzki, B.; Basler, K.; Ahrens, C.; Grossniklaus, U. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Res.* **2009**, *19*, 1786−1800.
(22) Gerster, S.; Qeli, E.; Ahrens, C.; Bühlmann, P. Protein and gene model inference based on statistical modeling in k-partite graphs. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 12101−12106.
(23) Görg, A.; Weiss, W.; Dunn, M. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004**, *4*, 3665−3685.
(24) Pedersen, S.; Harry, J.; Sebastian, L.; Baker, J.; Traini, M.; McCarthy, J.; Manoharan, A.; Wilkins, M.; Gooley, A.; Righetti, P.; Packer, N.; Williams, K.; Herbert, B. Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. *J. Proteome Res.* **2003**, *2*, 303−311.
(25) Prior, M.; Larance, M.; Lawrence, R.; Soul, J.; Humphrey, S.; Burchfield, J.; Kistler, C.; Davey, J.; La-Borde, P.; Buckley, M.; Kanazawa, H.; Parton, R.; Guilhaus, M.; James, D. Quantitative proteomic analysis of the adipocyte plasma membrane. *J. Proteome Res.* **2011**, *10* (11), 4970−4982.
(26) Gygi, S.; Corthals, G.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9390.
(27) 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.
(28) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R.; Olsen, J.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794−1805.
(29) Elias, J.; Gygi, S. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.
(30) Chang, C.; Lin, C. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, A27.
(31) Falcon, S.; Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **2007**, *23*, 257−258.
(32) Wu, C.; et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **2006**, *34*, D187−D191.
(33) Arnold, K.; Kiefer, F.; Kopp, J.; Battey, J.; Podvinec, M.; Westbrook, J.; Berman, H.; Bordoli, L.; Schwede, T. The protein model portal. *J. Struct. Funct. Genomics* **2009**, *10*, 1−8.
(34) Kiefer, F.; Arnold, K.; Künzli, M.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* **2009**, *37*, D387−D392.
(35) Sancak, Y.; Peterson, T.; Shaul, Y.; Lindquist, R.; Thoreen, C.; Bar-Peled, L.; Sabatini, D. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **2008**, *320*, 1496.
(36) Zoncu, R.; Efeyan, A.; Sabatini, D. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell Biol.* **2010**, *12*, 21−35.
(37) Schürmann, A.; Brauers, A.; Maßmann, S.; Becker, W.; Joost, H. Cloning of a novel family of mammlian GTP-binding proteins (RagA, $RagB^S$, $RagB^1$) with remote similarity to the ras-related GTPases. *J. Biol. Chem.* **1995**, *270*, 28982−28988.