Published in final edited form as:

J Proteome Res. 2007 September ; 6(9): 3549–3557. doi:10.1021/pr070230d.

Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency

Bing Zhang^{*,†}, Matthew C. Chambers[‡], and David L. Tabb^{†,‡,§}

Department of Biomedical Informatics, Mass Spectrometry Research Center, and Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8575

Abstract

Assembling peptides identified from LC–MS/MS spectra into a list of proteins is a critical step in analyzing shotgun proteomics data. As one peptide sequence can be mapped to multiple proteins in a database, naïve protein assembly can substantially overstate the number of proteins found in samples. We model the peptide–protein relationships in a bipartite graph and use efficient graph algorithms to identify protein clusters with shared peptides and to derive the minimal list of proteins. We test the effects of this parsimony analysis approach using MS/MS data sets generated from a defined human protein mixture, a yeast whole cell extract, and a human serum proteome after MARS column depletion. The results demonstrate that the bipartite parsimony technique not only simplifies protein lists but also improves the accuracy of protein identification. We use bipartite graphs for the visualization of the protein assembly results to render the parsimony analysis process transparent to users. Our approach also groups functionally related proteins together and improves the comprehensibility of the results. We have implemented the tool in the IDPicker package. The source code and binaries for this protein assembly pipeline are available under Mozilla Public License at the following URL: <http://www.mc.vanderbilt.edu/msrc/bioinformatics/>.

Keywords

parsimony analysis; bipartite graph; shotgun proteomics; LC-MS/MS; protein assembly

Introduction

Identification of all the proteins expressed in a cell or tissue is a fundamental task in proteomics. The liquid chromatography–tandem mass spectrometry (LC–MS/MS)–based proteomics approach has emerged as an effective strategy for protein identification in complex samples.¹ Because these experiments typically generate tens of thousands of tandem mass spectra, bioinformatics tools are necessary to derive biological information from the collected data. These bioinformatics tasks include duplicate spectrum recognition,^{2,3} peptide charge state discernment,⁴ peptide identification,^{5,6} protein assembly,^{7–9} identification error rate assessment,^{10,11} and sample comparison.¹²

*To whom correspondence should be addressed., bing.zhang@vanderbilt.edu; phone, 615-936-0090; fax, 615-936-1427.

[†]Department of Biomedical Informatics, Vanderbilt University Medical Center.

[‡]Mass Spectrometry Research Center, Vanderbilt University Medical Center.

[§]Department of Biochemistry, Vanderbilt University Medical Center.

Supporting Information Available: Figures of the greedy set cover algorithm and the database size effect on the sensitivity of peptide identification. This material is available free of charge via the Internet at <http://pubs.acs.org>.

The bioinformatics process most characteristic of LC-MS/MS proteomics is that of peptide identification. This process is usually carried out through search engines such as SEQUEST⁵ and Mascot.⁶ Peptides are enumerated from a protein sequence database, and the fragment ions predicted for each are compared to the spectra. Because only a fraction of spectra can be matched successfully to peptide sequences, the raw identifications must be filtered to retain only those most likely to be accurate.

Protein assembly plays a critical role in identification as it transforms a list of identified peptides into a list of identified proteins. If each peptide was unique to a particular protein, assembly would simply be a matter of grouping peptides by the proteins in which they were found. Because peptides are frequently part of multiple proteins, however, this naïve protein assembly can substantially overstate the number of proteins found in samples. Sequence redundancy can result from a several causes. Gene duplication and subsequent mutation may result in sets of genes with similar sequences. Alternative splicing of a gene's hnRNA may be represented by protein isoforms that are listed as separate proteins in a database. Some sequence databases contain multiple sequences for proteins that represent alternative forms of proteins that differ by minor mutations. In some cases, databases give fragmentary forms of protein sequences along with full-length sequences. These redundancies are not lapses in curation; the biological reality that they represent is complex, and the rules used for constructing the sequence databases are complex in response.

Several bioinformatics tools have been developed to address the difficulties in protein assembly. DTASelect⁸ groups together proteins with identical sets of identified peptides and uses a similarity score to describe the relationship between proteins with overlapped peptide identifications. In addition, it can remove proteins for which the observed evidence is a subset of the peptides observed for another protein. DBParser⁹ has pushed forward along this line by classifying and reporting proteins in six hierarchical categories: distinct, differentiable, subsumable, superset, subset, and equivalent. Parsimony analysis is employed to derive the minimal protein list sufficient to account for the observed peptide identifications. These tools help reduce sequence redundancy problems. However, they ignore the relative quality of passing peptide identifications. A statistical approach has been proposed by Nesvizhskii et al. to compute probabilities that proteins are present in a sample on the basis of estimated peptide identification probabilities.⁷ This approach apportions peptide identifications among all corresponding proteins and derives the minimal protein list using the expectation-maximization algorithm. As this is conceptually equivalent to the principle of parsimony used in DBParser, minimal protein lists generated by these two approaches are actually identical.⁹ The model relies on the probabilities that peptide assignments are correct, which can be hard to obtain, especially when a data set involves peptides identified using different search algorithms and different search databases as a result of collaborative studies.¹³ It has also been suggested that the quantitative information observed at the peptide and protein levels may be used to allocate shared peptides in creating a minimal protein list.¹⁴ Although the statistical model-based approach has shown promising results, many proteomics labs continue to operate without systems to produce minimal protein lists from sets of peptide identifications.

Parsimony algorithms can substantially reduce the number of proteins reported from experiments, but to our knowledge, there have been no evaluations of the quality of the generated lists of protein identifications. Some laboratories have avoided adopting such systems because of concerns that correct protein identifications would be erroneously filtered out. A second concern that users have for these systems is that they are algorithmically complex, making it difficult for users to discern why particular proteins have been retained or rejected. This challenge may delay publication of proteomics data at a time when journals are adopting guidelines which require that "when a single protein member of a multiprotein family has been singled out, the authors should explain how the other members of the group were ruled out".

¹⁵ As proteomics standardization progresses, authors are likely to be required to describe the complex many-to-many relationships between identified peptides and the proteins that potentially explain their appearance. These challenges only increase the importance of transparency in protein parsimony, and they expose the limitations of protein assembly tools that generate unstructured protein and peptide lists.

In this paper, we characterize IDPicker, an open-source protein assembly tool. Instead of relying on peptide identification score thresholds, we estimate False Discovery Rates (FDR) from reversed-sequence database search to control the quality of the peptide identifications. Efficient bipartite graph algorithms are used to cluster proteins by shared peptides and to derive the minimal list of proteins, enabling the software to scale up to millions of peptide identifications collected in large-scale studies. The software is demonstrated using three data sets, including a defined human protein mixture, a yeast whole cell extract, and a human serum proteome, each searched with three different sequence databases. Here, we explore the reduction of protein lists in parsimony processing as a function of the database employed in identification. We evaluate the accuracy achieved in protein filtering by standard peptide-count-based approaches and through parsimony processing. We also examine the visualization of protein-peptide relationships through bipartite graph images and in HTML. The algorithms and implementation of IDPicker are likely to be of use to many proteomics labs.

Materials and Methods

Data Sets

Three shotgun proteomics data sets are used in this study. All three data sets can be downloaded from the ProteomeCommons.org Tranche system using the hashes available at <http://www.mc.vanderbilt.edu/msrc/bioinformatics/>.

Data Set I: Sigma49

In the Mass Spectrometry Research Center at Vanderbilt University, a mixture of 49 human proteins from Sigma Aldrich (St. Louis, MO) was electrophoresed into a short 10% SDS-PAGE gel for cleanup. The unseparated proteins were reduced, alkylated with iodoacetamide, and digested in-gel with trypsin. Three replicate LC-MS/MS analyses were produced in a Thermo LTQ instrument (San Jose, CA), yielding approximately 10 900 tandem mass spectra each. Spectra were identified using MyriMatch¹⁶ version 1.0.366 in three separate searches against the Swiss-Prot database (241 365 sequences), a subset database containing the Swiss-Prot proteins with the “_HUMAN” identifier (15 638 sequences), and the IPI Human database, version 3.26 (67 665 sequences). In all three cases, the reversed version of each protein sequence was appended. Only tryptic peptides were considered. All cysteines were assumed to be carboxamidomethylated, and methionines were allowed to be oxidized. A precursor error of up to 1.25 m/z was permitted, while fragment ions were required to fall within 0.5 m/z of their expected locations. Ambiguous identifications that mapped spectra to multiple peptide sequences at equal scores were excluded.

Data Set II: Yeast-Extract

Students in Dr. Andrew Link's section of the 2006 Cold Spring Harbor Laboratory Proteomics course generated four replicate MudPits of yeast whole cell extracts. Each replicate, produced on a Thermo LTQ coupled with a microcapillary LC and nanospray ESI source, contained approximately 40 600 MS/MS in six SCX/RPLC separations. Spectra were identified similarly to above, employing the full Swiss-Prot sequence database, the “_YEAST” subset of Swiss-Prot (6090 sequences), and the *Saccharomyces* Genome “or-f_trans_all” Database (6714 sequences) for the searches. Of these, only the full Swiss-Prot database would include likely contaminants such as human keratins and porcine trypsin.

Data Set III: Serum-MARS

A human serum sample was processed through the Multiple Affinity Removal System (MARS) to deplete albumin, transferrin, IgG, IgA, anti-trypsin, and haptoglobin. Ten replicates were produced by the Western Consortium of the National Cancer Institute's Mouse Models of Human Cancer Consortium.¹⁷ Each replicate yielded approximately 23 500 tandem mass spectra. The data were downloaded from the ProteomeCommons Web site (<http://www.proteomecommons.org/data/annotations/860.html>). Spectra were identified against the Swiss-Prot database, the human subset of Swiss-Prot, and the IPI Human database, version 3.26.

IDPicker Implementation

IDPicker is a pipeline of tools designed to assemble confident, parsimonious protein identifications from raw spectral identifications. This software includes three modules, described in detail in the following sections. The first tool, written in C++, reads the unfiltered peptide identifications from a SQT file,¹⁸ applies an initial pass of filtering (typically to an FDR of 25%), and records plausible identifications with their FDR assessments to an XML file. The second tool groups these reported XML files into appropriate sets (e.g., grouping together MudPit identifications or defining technical replicates), filters peptides to the final FDR (typically 5%), and writes this information to a new XML file. The final tool applies parsimony analysis to discovered proteins and produces reports. This pipeline of tools is designed to work with identifications regardless of the database search tool that created them. Source code for the tools is available at <http://www.mc.vanderbilt.edu/msrc/bioinformatics/>.

Deriving Error Estimates for Peptide Identification

The software determines identification score thresholds that correspond to user-specified FDR. It examines the peptide identifications at a series of thresholds to determine the numbers of peptide identifications derived from the forward-sequence database (F) and those derived from the reverse-sequence database (R). The peptide matches to forward and reversed sequences are used to compute an FDR by this formula: $FDR = (2R)/(F + R)$. The FDR recorded for a particular peptide identification is the FDR of the set of identifications that would result if this identification were the lowest scoring identification to pass.

In the current implementation, all peptides that pass this initial filter are given equal standing. Identifications with higher scores are not given greater importance in protein assembly than those with lower scores. Likewise, the identifications below the threshold are removed entirely from consideration. Additional low-scoring peptides may be associated with a protein that is given in the final list, but if they exceed the required FDR, they are not reported. The Boolean nature of peptide inclusion in IDPicker may be extended in later revisions.

Bipartite Graph Analysis of the Peptide Identification Data

After deriving the list of peptides at a certain FDR level, we collect all proteins that could account for these peptides. This protein list represents the maximum possible number of proteins for this database search that could explain the observed spectra. The mapping between proteins and peptides is very complex, as one peptide could map to multiple proteins, and one protein could account for multiple peptides. This complex relationship can be modeled by a bipartite graph. In graph theory, a bipartite graph is an undirected graph where vertices can be partitioned into two sets such that no edge connects vertices in the same set. Intuitively, a bipartite graph is one whose vertices can be colored red and blue such that no edge exists between like colors. Figure 1 illustrates the bipartite graph analysis procedure, which includes four steps: initialize, collapse, separate, and reduce.

(1) Initialize—As shown in Figure 1A, we represent the peptide identification data in a bipartite graph that includes two sets of vertices: protein vertices (Figure 1A, pro1–pro9) and peptide vertices (Figure 1A, pep1–pep10). This graph structure only allows connections between peptide vertices and protein vertices. A peptide vertex and a protein vertex are connected by an edge if the peptide sequence could be mapped to the protein sequence.

(2) Collapse—Some protein vertices are connected to exactly the same set of peptide vertices (Figure 1A, pro2 and pro8; pro4 and pro9). These proteins are indiscernible from each other given the peptide identification results. We define a meta-protein as a group of indiscernible proteins based on available evidence. In Figure 1B, we collapse these indiscernible protein vertices together to form a meta-protein vertex (Figure 1B, pro2,8; pro4,9). Similarly, we collapse peptide vertices that connect to exactly the same set of protein vertices to form a meta-peptide vertex (Figure 1B, pep1,5; pep3,7,9). The final version of a bipartite graph has two sets of vertices: meta-protein vertices and meta-peptide vertices.

(3) Separate—Protein assembly is complicated by degenerate peptides that match to more than one protein. Two proteins are independent with regard to protein assembly if they share no peptides directly or indirectly through other proteins. Therefore, in Figure 1C, we seek to decompose the complex bipartite graph into independent subgraphs of proteins with shared peptides. We achieve this through the depth first search¹⁹ to enumerate connected components from a bipartite graph. A connected component is a maximal connected subgraph. Each connected component represents a meta-protein cluster, in which every meta-protein has shared meta-peptides with at least one other meta-protein in the same cluster.

(4) Reduce—To create a list of meta-proteins that are necessary for the explanation of the observed spectra, we generate a minimal list of meta-proteins for each meta-protein cluster as shown in Figure 1D. We achieve this through a greedy set cover algorithm¹⁹ to derive a minimum subset of meta-protein vertices that cover all of the peptide vertices in a protein cluster. The greedy algorithm iteratively chooses a meta-protein vertex that connects to the largest number of uncovered meta-peptide vertices in each stage. When all meta-peptides have been accounted for, the remaining meta-proteins are eliminated. Meta-proteins that have been selected as part of the set cover constitute the parsimonious protein list (highlighted in Figure 1D). The greedy set cover algorithm is illustrated in the Supporting Information (Figure S1).

Different Methods for Generating Protein Lists

As proteins in the same meta-protein are indistinguishable given the peptide sequences, all the analyses in this paper will be based on the protein identification at the meta-protein level. We compare four methods of generating the meta-protein list. PEP1 retains all meta-proteins even if their only evidence is a single shared peptide. PEP1-PARS applies the parsimony analysis on the PEP1 list to generate the minimal protein list. PEP2 retains only meta-proteins matching to at least two different peptide sequences. PEP2-PARS applies the parsimony analysis on the PEP2 list to generate the minimal protein list.

Results and Discussion

We used three data sets from very different experimental settings to test the effects of this parsimony software. The three data sets used in the study, *Sigma49*, *Yeast-Extract*, and *Serum-MARS* represent a defined human protein mixture, a yeast whole cell extract, and a human serum proteome, respectively. Because parsimony algorithms perform differently in response to sequence databases of different size and complexity, we employed three different databases for each sample: the multi-species database Swiss-Prot (SP), the comprehensive species-specific databases IPI Human (IPI) for human and *Saccharomyces* Genome Database (SGD)

for yeast, and compact species-specific subsets of Swiss-Prot (SPH for human and SPY for yeast).

Protein List Reduction

As requiring two peptides per protein is routine in many laboratories, we employed this filter in testing the impacts of grouping indiscernible proteins and parsimony analysis. Replicates in each data set for each database were averaged and plotted in Figure 2. All reported counts include only the proteins in the normal orientation; reversed proteins are excluded. The white bars count each protein separately whether it can be distinguished from others on the basis of observed peptides or not. The gray bars reflect the result of grouping indiscernible proteins into meta-proteins. The black bars show the number of these meta-proteins that remain after parsimony analysis. The protein counts were reduced most when Swiss-Prot was employed (Figure 2, SP groups). The large reductions reflect that Swiss-Prot contains many species, leading to artifactual identification of orthologous proteins.

The protein lists were less redundant when organism-specific databases were employed for the search, with human databases leaving more room for reduction than yeast databases. For the *Sigma49* and the *Serum-MARS* data sets (Figure 2, panels A and B), when the comprehensive human IPI database was employed for the search, the replicate protein counts were reduced by 39% and 24% by grouping indiscernible proteins. The parsimony analysis reduced the counts by an additional 51% and 44% in the two data sets. The reduction was relatively minor when the compact SPH database was employed; it contains considerably fewer duplication and isoform variations than the IPI database. Therefore, the protein counts reduced by only 8% and 5% in the *Sigma49* and the *Serum-MARS* data sets as indiscernible proteins were grouped. The parsimony analysis reduced the meta-protein list by 12% and 4% more in the two data sets.

The *Yeast-Extract* data set searched against the SGD represents a case in which little parsimony reduction would be expected. Grouping indistinguishable proteins into meta-proteins resulted in a 14% reduction, and parsimony analysis gave an additional 3% reduction, reflecting the minimal homology among sequences in the yeast proteome (Figure 2C). These reductions show that grouping indiscernible proteins and parsimony analysis can improve protein reporting even in organisms with compact genomes.

Grouping indiscernible proteins and parsimony analysis can help to abstract away differences in sequence database size. In the *Sigma49* runs, for example, the average initial numbers of proteins were 414 for SP, 161 for IPI, and 59 for SPH (white bars in Figure 2A). After the two reductions were applied, the meta-protein counts were 51 for SP, 49 for IPI, and 48 for SPH (black bars). The actual protein overlaps between the 51 SP, 49 IPI, and 48 SPH meta-proteins and the known proteins in the original *Sigma49* sample are 37, 32, and 40, respectively. As the true content of the sample was given as a list of Swiss-Prot IDs, we converted the IPI IDs to Swiss-Prot IDs based on the annotation provided by the IPI Web site. The relatively low overlap for the IPI meta-proteins may have resulted from the imperfect mapping between IPI IDs and the Swiss-Prot IDs. These list reduction strategies cause the resulting protein lists to converge to numbers far closer to the true number of proteins in the sample. Although using extremely large databases can reduce the sensitivity of peptide identification (see Figure S2 in the Supporting Information for the comparison), many laboratories routinely employ multispecies databases such as the NCBI nonredundant or EBI UniRef databases. In these cases, the proposed reduction method is extremely useful for removing the redundancy in the protein list.

Improved Accuracy of Protein Identification

Reducing the size of protein lists is useful only if incorrect protein identifiers are the ones being removed. We used the *Sigma49* and the *Yeast-extract* data sets to evaluate the effects of the bipartite graph modeling approach on the accuracy of protein identification.

The *Sigma49* data set was produced from a sample containing 49 defined human proteins. The protein components in the *Sigma49* sample were given as a list of Swiss-Prot IDs. To avoid the possible effect of imperfect ID mapping, we excluded the IPI search for this analysis. Each meta-protein in the final meta-protein list was counted as a true positive (TP) if it included one of the 49 proteins listed as part of the sample or as a false positive (FP) otherwise. We compared the four methods for generating protein lists as described in Materials and Methods for their performance using the metrics developed for the evaluation of information retrieval systems, including precision, recall, and F_1 -measure.²⁰

Precision is defined as the proportion of identified relevant items to all identified items, and can be calculated as $n_{TP}/(n_{TP} + n_{FP})$, where n_{TP} is the number of true positives and n_{FP} is the numbers of false positives. Recall is defined as the proportion of identified relevant items to all relevant items, and can be calculated as n_{TP}/n_P , where n_P is the number of all proteins in the sample, that is, 49 in this analysis. F_1 -measure is the harmonic mean of recall and precision, and can be calculated as $2pr/(p + r)$, where p is the precision and r is the recall of the identification. For each method, we calculated the recalls, precisions, and F_1 -measures for the three replicates separately. The results from the replicates were averaged for presentation in Figure 3.

For the Swiss-Prot (SP) search (Figure 3A), the PEP1 method (white bars) gave a high recall of 0.89. However, it falsely identified more than 160 meta-proteins in each replicate, which led to a low precision of 0.20 and a low F_1 -measure of 0.33. Applying the parsimony analysis (PEP1-PARS: light gray bars) did not remove any true protein identifications in the course of eliminating many false identifications and increased the precision and F_1 -measure to 0.55 and 0.68, respectively. Applying the PEP2 filtering alone (dark gray bars) yielded improved precision, but it also reduced the recall, and as a result, the F_1 -measure only improved slightly compared to PEP1. When we combined the parsimony analysis and the two-peptide filtering in PEP2-PARS (black bars), we achieved the highest F_1 -measure of 0.74, which is 2.2 times that of PEP1. These results demonstrate that the bipartite graph approach is highly effective at removing false protein identifications while retaining true identifications. Coupling the parsimony analysis with the two-peptide filtering can further reduce false identifications and improve the overall accuracy of protein identification.

Searching the compact Swiss-Prot human subset (SPH) would not seem to leave much room for improvement for parsimony analysis. However, the parsimony analysis alone (PEP1-PARS) improved the F_1 -measure by 9%, and the combination of the parsimony analysis and the two-peptide filtering (PEP2-PARS) improved the F_1 -measure by 23% (Figure 3B). Unlike the SP search, in which the parsimony analysis was much more effective than the two-peptide filtering with regard to the F_1 -measure improvement, in the SPH search, the parsimony analysis was less effective. This suggests that the parsimony analysis is most useful in removing redundant homologous proteins. Therefore, it will be most powerful in processing data sets generated by searching multispecies databases such as the Swiss-Prot, or comprehensive single-species databases for higher organisms such as the IPI Human.

Because the protein content in the *Yeast-Extract* data set was not defined, it is impossible to calculate the F_1 -measure. The Swiss-Prot database used for the SP search has 5916 Yeast proteins and 235 449 proteins from other organisms. We used the SP search result to evaluate the performance of parsimony analysis by its specificity in removing non-yeast proteins. We

made a rough estimation by treating each meta-protein in the final meta-protein list as a true positive (TP) if it included at least one protein with the “_YEAST” identifier, or a false positive (FP) if it did not. For each filtering strategy described in Materials and Methods, we calculated the number of TPs (n_{TP}) and FPs (n_{FP}) for the three replicates separately. The results from the replicates were averaged for presentation in Figure 4. A good protein identification method should keep as many true positives as possible while eliminating false positives and, thus, lead to a high n_{TP}/n_{FP} ratio. PEP1 gave an n_{TP}/n_{FP} ratio of 1.4. Applying the parsimony analysis in PEP1-PARS doubled the ratio to 2.8, and only 3.5% of the true positives were lost. Applying the two-peptide filtering alone (PEP2) increased the ratio to 2.4, but it reduced n_{TP} by 40% compared to PEP1. When we combined the parsimony analysis and the two-peptide filtering in PEP2-PARS, we achieved the highest n_{TP}/n_{FP} ratio of 13.4, reducing n_{TP} by only 6% more than PEP2. It is clear that the parsimony analysis significantly increased the n_{TP}/n_{FP} ratio with or without two-peptide filtering. Using PEP2 filtering lost a considerable number of true identifications in the *Yeast-Extract* data set. However, some of the removed yeast proteins could actually be false identifications, as we obviously overestimated the number of true positives. The results from yeast demonstrate that the bipartite graph parsimony approach is effective in screening out orthologous protein hits while retaining proteins from the correct species.

Increased Transparency and Functional Clustering for the Visualization of Protein Identification Results

The bipartite parsimony approach simplifies protein lists, removes spurious protein identifications, and provides a natural way for visualizing the results. The IPI search results on the *Serum-MARS* data set reveal several aspects of this approach that recommend it for the practice of proteomics. We focused on the PEP2 and PEP2-PARS methods for all 10 replicates in aggregate. The 10 LC separations produced a total of 194 523 tandem mass spectra. Because spectra from multiply charged peptides were identified under two different charge state assumptions, 350 648 identifications resulted from the database search. IDPicker filtered these identifications down to 37 246 to achieve a 5% FDR for identifications. Many of these identifications were duplicates or different charge state variants of the same peptide, and the software found that 2605 different peptide sequences were represented. These 2605 peptides could be explained by as many as 472 proteins (including reversed sequences), but these could be reduced to 339 distinguishable meta-proteins and subsequently to 189 meta-proteins after parsimony analysis. After database searching, the entire analysis was processed on a single CPU in minutes, with the slowest part being the reading and filtering of raw identifications. The software has been designed to scale for data sets that encompass tens of millions of raw identifications.

Users of parsimony algorithms need the ability to trace the fates of particular proteins. In IDPicker, one can see the interrelationships of proteins with or without the parsimony option in effect. Figure 5 shows the result for haptoglobins in the aggregate analysis of the *Serum-MARS* data set. Without parsimony, five proteins are clustered together to explain an overlapping set of peptides. IDPicker produces a tabular list of the proteins (see Figure 5A), association tables revealing which meta-proteins (rows) map to which meta-peptides (columns, see Figure 5B,C), and a graphic illustrating the relationship among the five proteins and seven meta-peptides (see Figure 5D). Users can expand a table enumerating the peptides and identifications that are part of each meta-peptide (not shown). Peptides are called different sequences if they differ by particular residues or by post-translational modifications.

The second meta-peptide of Figure 5B contains three different sequences that were identified in 17 different spectra. This meta-peptide appears as the second block of peptides in Figure 5D. The two edges leading to proteins in this figure match to the two “X” symbols in the column

of Figure 5B for this meta-peptide; the association tables show the same information as the image. The number of sequences and spectra associated with a particular protein in 5A sums across all meta-peptides that match to the protein. The second meta-peptide of Figure 5B contributes to the counts for both IPI00477597.1 and IPI00607707.1 (meta-proteins C and D). After parsimony has been applied, the second meta-peptide of Figure 5B can be merged with the first because meta-protein D has been removed, and the first meta-peptide of Figure 5C reflects the merger of this pair. Routine use of parsimony filtering can simplify peptide–protein relationships significantly.

When parsimony analysis is applied, the software finds that two meta-proteins can explain the presence of all observed peptides, and the reduced association table reflects the peptides corresponding to each (Figure 5B is reduced to Figure 5C; see the shaded proteins in Figure 5D). These visualizations enable users to determine why a given protein was removed from the list. They also enable the discernment of which peptides are shared among or distinct to proteins remaining in the list, an important requirement of the MIAPE guidelines²¹ (<http://psidev.sourceforge.net/miape/>). Parsimony analysis is conservative by nature and may remove biologically meaningful protein identifications. The transparency offered by the bipartite graph provides the users the possibility to use their judgment to modify the list generated by the computer, which is not possible if the parsimony analysis process is hidden from the users.

The clustering applied for the bipartite graph model has useful side-effects for the report of protein lists. We sorted the clusters in the *Serum-MARS* sample by the number of meta-proteins they contained. Below, we report the number of proteins in each cluster both with and without parsimony applied for the top five clusters (where $X \rightarrow Y$ indicates that X proteins were present in the cluster prior to parsimony and Y afterward):

- 11 $\rightarrow$ 8 complement factor H-related proteins
- 30 $\rightarrow$ 4 type I keratins
- 8 $\rightarrow$ 4 plasminogens and lipoproteins
- 11 $\rightarrow$ 3 type II keratins
- 6 $\rightarrow$ 2 complement C3 proteins

Clustering proteins by their shared peptides has the helpful effect of grouping them by sequence similarity and therefore functional similarity. Clustering proteins based on sequence similarity has been used in Mascot's Integra (<http://www.matrixscience.com/integra.html>) for reporting identified proteins. Bundling large families of proteins such as keratins into single clusters instead of having them scattered throughout the protein list makes human serum data easier to review. Our ability to find features of interest in proteomics data sets has been greatly accelerated by this clustering. In another cluster, one of the proteins was annotated as “97 kDa protein”. Because its partner in the cluster was annotated as “ceruloplasmin precursor,” we could get a sense of its function far more rapidly than if each protein was listed separately.

Conclusions

The bipartite graph is a useful model for representing peptide identification data in LC–MS/MS proteomics. It provides efficiency, accuracy, and transparency in deriving a minimal protein list from peptide identifications. Fast graph algorithms make it possible to process hundreds of thousands of peptide identifications on a single CPU in minutes. These algorithms have been implemented in the IDPicker software, which is scalable for data sets that encompass tens of millions of raw identifications. Applying the IDPicker software on three shotgun proteomics data sets demonstrated substantial protein list reduction, especially for human

samples. As shown in the *Sigma49* example, parsimony processing reduced the protein lists to a size that more closely resembled the true number of proteins identifiable in a sample, irrespective of sequence database size. In both the *Sigma49* and *Yeast-Extract* sets, the bipartite graph analysis was highly efficient in removing false protein identifications while retaining true identifications. This effect was complementary to the specificity improvements observed by requiring more peptides per protein. The bipartite graph model also yields an intuitive visualization of protein–peptide relationships for improved transparency of this process for end users. Moreover, it groups functionally related proteins together through clustering proteins with shared sequences and, thus, helps users to examine results more efficiently. The bipartite graph approach to protein parsimony has much to offer in standardizing protein assembly and parsimony processing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH/NCI 1 R01 CA126218-01, NIH/NCI 1 U24 CA126479-01, and NIH P30 ES000267. Dr. Amy Ham and Ms. Kristin Cheek in the Proteomics Laboratory of the Vanderbilt Mass Spectrometry Research Center produced the mass spectrometry data from the *Sigma49* samples. Dr. Andrew Link at Vanderbilt School of Medicine provided the *Yeast-Extract* data from the 2006 CSHL Proteomics course. The CSHL Proteomics course was supported by Grant Number 2T15 CA098595 from the National Cancer Institute. We thank Amanda Paulovich et al. for making the Serum-MARS set publicly available under NCI contract 23XS144A.

References

1. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;19(3):242–247. [PubMed: 11231557]
2. Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004;4(4):950–960. [PubMed: 15048977]
3. Tabb DL, Thompson MR, Khalsa-Moyers G, VerBerkmoes NC, McDonald WH. MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J Am Soc Mass Spectrom* 2005;16(8):1250–1261. [PubMed: 15979332]
4. Klammer AA, Wu CC, MacCoss MJ, Noble WS. Peptide charge state determination for low-resolution tandem mass spectra. *Proc IEEE Comput Syst Bioinform Conf* 2005:175–185. [PubMed: 16447975]
5. Eng JK, McCormack AL, Yates JR 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. Perkins DN, Pappin DJ, Creasy DM, Cottrell JC. Probability based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
7. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75(17):4646–4658. [PubMed: 14632076]
8. Tabb DL, McDonald WH, Yates JR III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 2002;1(1):21–26. [PubMed: 12643522]
9. Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, Chen X, Markey SP, Kowalak JA. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* 2004;3(5):1002–1008. [PubMed: 15473689]
10. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4(3):207–214. [PubMed: 17327847]
11. Keller A, Nesvizhskii AI, 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(20):5383–5392. [PubMed: 12403597]

12. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 2006;5(11):2909–2918. [PubMed: 17081042]
13. Adamski M, Blackwell T, Menon R, Martens L, Hermjakob H, Taylor C, Omenn GS, States DJ. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* 2005;5(13):3246–3261. [PubMed: 16104057]
14. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005;4(10):1419–1440. [PubMed: 16009968]
15. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* 2004;3(6):531–533. [PubMed: 15075378]
16. Tabb DL, Fernando CG, Chambers MC. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 2007;6(2):654–661. [PubMed: 17269722]
17. Whiteaker JR, Zhang H, Eng JK, Fang R, Piening BD, Feng LC, Lorentzen TD, Schoenherr RM, Keane JF, Holzman T, Fitzgibbon M, Lin C, Zhang H, Cooke K, Liu T, Camp DG II, Anderson L, Watts J, Smith RD, McIntosh MW, Paulovich AG. Head-to-head comparison of serum fractionation techniques. *J Proteome Res* 2007;6(2):828–836. [PubMed: 17269739]
18. McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, Johnson JR, Cociorva D, Yates JR III. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* 2004;18(18):2162–2168. [PubMed: 15317041]
19. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. *Introduction to Algorithms*. 2. MIT Press and McGraw-Hill; Cambridge. MA: 2001.
20. van Rijsbergen, CJ. *Information Retrieval*. 2. Butterworths; London: 1979.
21. Taylor CF. Minimum reporting requirements for proteomics: a MIAPE primer. *Proteomics* 2006;6 (Suppl 2):39–44. [PubMed: 17031795]

Figure 1.

Overview of the bipartite graph model and analysis procedure for protein assembly. (A) Initialize: connect peptides to all proteins that contain them in a bipartite graph. (B) Collapse: collapse indistinguishable protein and peptide vertices together to form meta-protein and meta-peptide vertices. (C) Separate: decompose the graph into protein clusters with shared peptides. (D) Reduce: identify and highlight a minimum protein list.

A. Sigma49 **B. Serum-MARS** **C. Yeast-Extract** **Figure 2.**

Reduction of the protein identification counts through the bipartite graph modeling approach. (A) The *Sigma49* data set; (B) the *Serum-MARS* data set; (C) the *Yeast-extract* data set. The white, gray, and black bars represent the initial protein counts from the database search, the meta-protein counts after grouping the indistinguishable proteins, and the minimal meta-protein counts after the parsimony analysis, respectively. The error bars represent the standard deviation from replicates. SPH, IPI, SP, SPY, and SGD represent different databases employed for the searches. SPH, Swiss-Prot human subset; SPY, Swiss-Prot yeast subset; IPI, IPI Human; SP, Swiss-Prot; SGD, *Saccharomyces* Genome Database.

Figure 3.

Recall, precision, and F_1 -measure for different methods of generating protein lists from the *Sigma49* data set. (A) Results from the Swiss-Prot search; (B) results from searching the Swiss-Prot human subset. The white, light gray, dark gray, and black bars represent the following four methods respectively: PEP1, PEP1-PARS, PEP2, and PEP2-PARS. PEP1 retains all meta-proteins. PEP1-PARS applies the parsimony analysis on the PEP1 list to generate the minimal protein list. PEP2 retains only meta-proteins matching to at least two different peptide sequences. PEP2-PARS applies the parsimony analysis on the PEP2 list to generate the minimal protein list. The error bars represent the standard deviation from three replicates.

Figure 4.

True and false protein identification counts for different methods of generating protein lists from the *Yeast-Extract* data set. The white and gray bars represent true positives (TP) and false positives (FP), respectively. PEP1, PEP1-PARS, PEP2, and PEP2-PARS represent four different methods for generating protein lists. PEP1 retains all meta-proteins. PEP1-PARS applies the parsimony analysis on the PEP1 list to generate the minimal protein list. PEP2 retains only meta-proteins matching to at least two different peptide sequences. PEP2-PARS applies the parsimony analysis on the PEP2 list to generate the minimal protein list. The error bars represent the standard deviation from four replicates.

A: Maximal protein list

28 total sequences, 386 total spectra

GID	Protein	Sequences	Spectra	Description
A	IPI00641737.1	24	368	Haptoglobin precursor
B	IPI00478493.3	20	339	HP protein
C	IPI00477597.1	19	213	Isoform 1 of Haptoglobin-related protein precursor
D	IPI00607707.1	18	212	Isoform 2 of Haptoglobin-related protein precursor
E	IPI00431645.1	14	243	HP protein

B: Protein/peptide association table

	1	2	3	4	5	6	7
Sequences	1	3	2	7	4	4	7
Spectra	1	17	34	139	29	62	104
A			x	x	x	x	x
B				x	x		x
C	x	x			x	x	x
D		x			x	x	x
E				x			x

C: Parsimonious association table

	1	2	3
Sequences	4	9	15
Spectra	18	173	195
A		x	x
C	x		x

D: Bipartite graph **Figure 5.**

Parsimony report generated by IDPicker. (A) Maximal protein list; (B) protein–peptide association table; (C) parsimonious association table; (D) bipartite graph visualization, proteins in the minimal list are shaded in gray.