THE EFECTS OF PARSIMONY LOGIC AND EXTENDED PARSIMONY CLUSTERING ON PROTEIN IDENTIFICATION AND QUANTIFICATION IN SHOTGUN PROTEOMICS

By

Raviteja Madhira

A THESIS

Presented to the Department of Medical Informatics & Clinical Epidemiology
and the Oregon Health and Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

December 2016

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Raviteja Madhira

"The Effects of Parsimony Logic and Extended Parsimony Clustering on Protein Identification and Quantification in Shotgun Proteomics"

Has been approved

Dr. Aaron Cohen

Dr. Phillip Wilmarth

Dr. Arie Baratt

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Peter Kurre (OHSU) for permission to use the mouse dataset in this work. I want to thank my thesis committee, Dr. Aaron Cohen (Thesis Mentor), Dr. Phillip Wilmarth (Research Mentor), and Dr. Arie Baratt (Committee Member) for their invaluable guidance on my project. I feel lucky to have a committee that helped me gain vast domain knowledge in proteomics, instilled ideas through productive discussions, and assisted me during any difficulties through the entire thesis duration.

Abstract

Shotgun proteomics is the most common mass spectrometry-based proteomics method for identifying and quantifying proteins present within a sample. Despite improvements in mass spectrometry tools, the issue dealing with inferring proteins and quantifying them from peptides still persists. The choice of protein sequence databases and the underlying genomic complexities of organisms are often considered sources of peptide degeneracy that lead to varying protein identifications and quantifications. In this thesis, the differences between four protein sequence database sources (Uniprot, NCBI, Ensembl, and IPI) were compared using redundant and non-redundant protein counts and shared and unique peptide counts for two higher eukaryotic organisms (human and mouse) and a representative lower eukaryotic organism (yeast). It was also demonstrated that basic parsimony logic in the protein inference process yields protein and peptide identifications in real biological samples of higher eukaryotic organisms that are dependent on the protein sequence database of choice. To address parsimony logic shortcomings, two versions of extended parsimony clustering algorithms (Proteomic Analysis Workflow (PAW) clustering and Scaffold-like clustering) that group proteins with highly significant shared peptide evidence but low unique peptide evidence were implemented and tested. For human samples, these extended parsimony clustering algorithms have significantly reduced both mean shared peptide proportions across databases compared to that of basic parsimony logic, and produced protein identification numbers that are largely independent of protein sequence database choice. Few differences in protein and peptide characteristics were observed for yeast samples before or after implementing PAW or Scaffold-like clustering algorithms. Silhouette scores and gene enrichment analysis on the clusters of the extended parsimony clustering algorithms demonstrated that they are biologically and functionally coherent. From a quantitative perspective, there was a significant increase in mean quantitative information content (QIC) in human samples after PAW or Scaffold-like clustering compared to QIC computed after basic parsimony logic. The variation in the QIC values of human samples significantly decreased across databases after implementation of PAW or Scaffold-like algorithms.

KEYWORDS: Shotgun Proteomics, Protein Sequence Databases, Redundant, Unique, Shared, Proteomics Analysis Workflow Clustering, Scaffold-like Clustering, Quantitative Information Content

Introduction

1.1. Shotgun Proteomics

The proteome is the entire set of proteins produced or modified by an organism or a system, and proteomics is the large-scale study of those proteins (Anderson and Anderson 1996; Graves and Haystead 2002; Pandey and Mann 2000). In less than two decades, proteomics has matured into a powerful arsenal of tools capable of identifying and quantifying thousands of proteins in biological samples (Gygi and Aebersold 2000; Hebert et al. 2014; Keshishian et al. 2015; Pandey and Mann 2000). Proteomic experiments are driven by continual advances in biological mass spectrometry, computational biology, and bioinformatics (Aebersold and Mann 2003; Gygi et al. 2000; Makarov et al. 2006; Mann et al. 2001; Senko et al. 2013; Yates 2000).

Most proteomic experiments do not perform mass spectrometry directly on intact proteins (known as top-down proteomics (Kelleher 2004; Yates and Kelleher 2013)), but instead perform analyses of peptides produced by proteolytic digestions of proteins to characterize the biological samples. This bottom-up methodology is also known as shotgun proteomics (McDonald and Yates 2002; Yates et al. 2009; Zhang et al. 2013). In a typical shotgun proteomics experiment, enzymatic digestion of a protein mixture, usually with trypsin, generates an extremely complex peptide mixture.

Elaborate separation techniques are used in conjunction with electrospray tandem mass spectrometry (MS/MS) to identify individual peptide sequences. Peptide spectral matches (PSMs) are obtained from MS/MS spectra by comparison to peptide sequences generated from *in silico* digestions of protein databases available from several sources. The scoring algorithms that assign the most likely peptides to the spectra are known as search engines (Baldwin 2004; Domon and Aebersold 2006; Matthiesen 2007). Overall, the goal of shotgun proteomics is to utilize the peptide information obtained from MS/MS to infer the proteins present within the sample (Huang et al. 2012; Nesvizhskii et al. 2003; Nesvizhskii and Aebersold 2005; Ma et al. 2012).

Shotgun proteomics methods are widely used because peptide mixtures are easier to process chemically (typically by liquid chromatography) and to sequence using MS/MS than intact proteins. Advances in mass spectrometry instrumentation have dramatically improved the accuracy of peptides mass measurements, increased sensitivities for detecting low-abundance peptides by several orders of magnitudes, and sped up the sequencing of peptides such that complete proteomes can be measured in only hours (Hebert et al. 2014; Makarov et al. 2006; Senko et al. 2013).

However, a major challenge associated with shotgun proteomics has not improved; namely, the difficulty in inferring proteins from their constituent peptides (Nesvizhskii and Aebersold 2005). Upon the digestion of the sample into peptides, it is hard to associate peptides to its parent proteins. Peptides liberated during digestion that can arise from multiple proteins are called degenerate or shared peptides. In contrast to unique peptides that provide immediate and direct information about their parent proteins, shared peptides create ambiguities in determining their respective parent proteins (Nesvizhskii and Aebersold 2005; Nesvizhskii 2007; Duncan et al. 2010).

1.2. Protein Inference

Determining the likely proteins present in a complex biological sample from a collection of partially accurate peptide sequences is challenging. Advances in instrumentation have improved the quality of mass spectrometry data and increased the numbers of sequence assignments. Data analysis advances have allowed more precise estimates of peptide sequences errors and the development of methods for reducing those errors to manageable levels (Elias and Gygi 2007; Keller et al. 2002; Nesvizhskii 2010).

The general logic for protein inference is well described in Nesvizhskii and Aerbersold 2005 and can take many algorithmic forms (Huang et al. 2012; Nesvizhskii et al. 2003; Serang et al. 2010; Zhang et al. 2007). Each peptide sequence assigned to a tandem mass spectrum by a search engine will have one or more associated proteins (typically labeled with the protein database accessions). Peptides associated with single protein labels are called unique (here with respect to the protein database used in the search) peptides. The peptides that have multi-protein labels (potentially coming from more than one protein) are known as shared peptides. Figure 1.1 illustrates how proteins can be either distinguishable or indistinguishable depending on unique and shared peptide evidence.

The assignment of protein labels to peptide sequences by search engines allows the set of assigned peptides to be determined for each protein in the protein database. Protein inference is basically a set cover problem (https://en.wikipedia.org/wiki/Set_cover_problem) where the goal is to determine the minimal number of peptide sets (the proteins) that can explain all observed peptides (the confidently identified PSMs). This minimal list is commonly referred to as a parsimonious protein list (a common method to generate a parsimonious list is to remove proteins that are subsets of other proteins based on peptide evidence) and has been a *de facto* proteomics reporting standard since 2005 (Bradshaw et al. 2006). It is important to recognize that parsimonious protein lists are most often accompanied by a redefinition of unique and shared peptides. Before protein inference, shared and unique peptide status are defined with respect to the protein sequence database. Typically, after protein inference, shared and unique peptide status are then defined with respect to the parsimonious protein list (see Fig. 1.1).

While instrument and analysis advances have improved many aspects of protein inference, some factors are beyond the experimental data. In higher eukaryotic organisms such as humans and mice, single-nucleotide polymorphisms, RNA editing, alternate splicing, and post-translational modifications can yield similar protein products (Rappsilber and Mann 2002) making protein inference more difficult. The majority of genes in higher organisms have multiple introns in contrast to lower genomes like yeast where introns are rare. Advances in genomic sequencing technologies have dramatically increased the number of available protein sequence databases for a wide variety of organisms. There are now several large repositories of genomic and protein sequences, and they can vary considerably in the completeness and complexity of their available sequences.

1.3. Protein Databases

The majority of proteomics experiments are shotgun proteomics and nearly all bottom-up proteomics analyses use search engines, such as MASCOT (Perkins et al. 1999), X!Tandem (Craig and Beavis 2004), and SEQUEST (Eng et al. 1994), in conjunction with protein databases to characterize biological samples. Despite the central role that protein databases play in proteomics, there is little consensus on database choices, particularly for important research organisms such as human and mouse. There are several organizations that produce protein sequence databases suitable for proteomics use such as the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov), the Universal Protein Resource (Uniprot, http://www.uniprot.org), the European Bioinformatics Institute and Wellcome

Trust Sanger Institute joint project (Ensembl, www.ensembl.org), and the International Protein Index that was maintained by the European Bioinformatics Institute (EBI) and Ensembl (IPI, www.ebi.ac.uk/ipi).

The NCBI database is an extremely large sequence repository with translations of nucleotide sequences from DNA Data Bank of Japan (DDBJ), EMBL Data Library, and GenBank databases. To manage the large number of sequences and facilitate biological research, protein sequences are processed to create a smaller set of Reference Sequences (Pruit et al. 2005) (RefSeq, http://www.ncbi.nlm.nih.gov/refseq/about/). RefSeq database records are curated, non-redundant, and explicitly link genomic, transcript, and protein sequences. Due to a good balance of completeness, redundancy, and high quality of sequence annotations, RefSeq protein databases are often used for protein/peptide identification in shotgun proteomic studies.

UniProt (Bairoch et al. 2005) is a consortium between the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI), and the Protein Information Resource (PIR). Uniprot is primarily focused on maintaining high-quality protein sequence databases. The UniProt Knowledgebase is composed of two protein database sections: Swiss-Prot and TrEMBL. The Swiss-Prot database is a manually curated, non-redundant database that includes rich annotations for all of its sequences. The curator selects one canonical protein sequence to represent each gene and any alternative protein forms (isoforms) are annotated as differences with respect to the canonical sequence. Protein sequences from UniProt may have large numbers of annotated isoforms for some species (for Swiss-Prot entries only), and protein databases can contain just the canonical Swiss-Prot sequences, or contain both canonical and isoform sequences.

The available protein annotations for Swiss-Prot include properties such as functions of proteins, Gene Ontology (GO) terms, post-translational modifications, domains, secondary and quaternary structures, similarities to other proteins, pathways of proteins, sequence conflicts, and cross-references to many other biological databases. Because of its high quality protein annotations, Swiss-Prot databases are not only used widely in proteomic studies, but also for understanding the functional and biological properties of proteins (Boeckmann et al. 2005).

Because such a highly curated database is labor-intensive and time consuming to generate, the TrEMBL (Translation to EMBL) database was created to automatically annotate proteins and incorporate new protein sequences more quickly (Apweiler et al. 2004; Boeckmann et al. 2003). Proteins in this database don’t have the rich annotations found in Swiss-Prot, but (in combination with Swiss-Prot) they help ensure completeness and maintain low levels of redundancy by merging records that contain identical full-length protein sequences for the same gene.

Ensembl (Aken et al. 2016; Flicek et al. 2012) is a genomic oriented database that strives for accurate translations of DNA sequences to protein sequences to produce species-specific reference proteomes. Ensembl provides a large number of reference genomes for higher eukaryotes, and their genome assemblies are frequently used as references in alignments of next generation sequencing data. There are excellent resources for comparative genomics (Herrero et al. 2016) and the BioMart tool offers many options for additional annotations. In investigations where both transcriptomics and proteomics are being studied, databases from the same source, such as Ensembl, can facilitate results comparisons.

IPI (Kersey et al. 2004), in use from 2001 to 2011, was a widely used protein sequence database resource for mass-spectrometry studies. Although IPI only had protein sequences of seven model organisms (human, mouse, rat, cow, chicken, zebrafish, and Arabidopsis), they were considered as having a good balance between completeness and degree of redundancy. The final release is still available for download; however, Reference Proteomes from UniProt are the recommended replacement protein databases (Griss et al. 2011). While protein annotations were somewhat minimal, it did maintain many cross-references with other protein databases such as RefSeq, Ensembl, and Uniprot (which could be out-of-date in many cases).

1.4. Quantitative Proteomics

In addition to protein identification, shotgun proteomics is also widely used for protein expression studies using methods such as label-free quantification, metabolic labeling, and chemical labeling where relative protein expression levels are inferred from relative peptide expression levels (Bantscheff et al. 2007; Bantscheff et al. 2012; Gygi et al. 2000; Ong and Mann 2005; Zhang et al. 2013). The ambiguity of shared peptides for protein inference also leads to ambiguity in their quantitative information content.

One common solution to shared peptides in quantitative proteomics is to completely discard them and use only the unique peptides for quantification. Figure 7 in Nesvizhskii and Aerbersold 2005 suggests that tryptic peptide degeneracy may vary considerably between different protein database choices and generally increases with larger, more complete databases. Higher eukaryotic organisms have genomic structure that may result in multiple distinct proteins that contain many shared tryptic peptides. Database sources described above have different goals. Some strive for completeness; others, such as Swiss-Prot, use manual curation to reduce sequence complexity and provide more precise biological contexts. It is likely that peptide degeneracy will vary depending on protein database choice and this may influence the number of usable (unique) peptides available for quantitation.

Figure 1.1 shows how definitions of shared and unique peptides can be context dependent and that they change during protein inference. Thus, the quantitative information content (QIC) (basically the fraction of unique peptides) may likely change depending on the context in which shared and unique peptides are defined. Parsimonious, or a minimal protein results that represent all the observed peptides, have become the accepted way to report proteomics results for publication, but no studies exploring how this process affects quantitative proteomics have been published.

The quality of sequencing data and the performance of mass spectrometers has improved dramatically since the original guidelines for parsimonious protein lists were laid out a decade ago (Bradshaw et al. 2006). Modern proteomics experiments are capable of producing millions of MS/MS spectra in a single study. It is possible that the basic parsimony logic may be overwhelmed by current large-scale proteomics datasets. Recent work (Koskinen et al. 2011) suggests that additional processing of parsimonious protein lists may be necessary to achieve a truly parsimonious set of identified proteins. Additional processing may add another context that changes definitions of shared and unique peptides and add another factor to consider in quantitative proteomic studies.

Methods and Datasets

2.1. Protein Databases

Protein sequences of human (taxon 9606), mouse (taxon 10090), and yeast (taxon 559292) were obtained in FASTA format (https://en.wikipedia.org/wiki/FASTA_format) from Uniprot, NCBI, Ensembl, and IPI sources using download options listed in Table 2.1. UniProt offers many options for protein sequence database downloads. As mentioned in the introduction, there are two mutually exclusive sections to UniProt, namely, Swiss-Prot (manually curated) and TrEMBL (computer annotated). Swiss-Prot has what could be considered as reasonably complete protein databases for only a small number of species, including human, mouse, and yeast. TrEMBL can best be thought of as a buffer of protein sequences that have not yet been added to Swiss-Prot. TrEMBL sequences should never be used without their companion Swiss-Prot sequences for a given organism. Since the retirement of IPI, UniProt has implemented Reference Proteomes. These are combinations of Swiss-Prot and a more curated set of TrEMBL sequences for a large number of species. Sequences in TrEMBL that lack experimental evidence for existence and that are fragments of other sequences seem to be filtered out, although processing details are difficult to ascertain.

Human and mouse are among the species for which large numbers of isoforms are annotated and available. Since UniProt approaches isoforms from an annotation perspective, isoforms are associated with Swiss-Prot sequences. Thus Swiss-Prot sequences are available as canonical sequences only or canonical sequences with known isoforms. This effectively doubles the number of protein database choices for each species of interest. Isoforms are represented in TrEMBL as distinct entries until they get removed and combined into single Swiss-Prot records by curators. It is not clear if databases for species like human or mouse, where large numbers of isoforms have been removed from TrEMBL, can be considered complete without inclusion of Swiss-Prot isoforms.

NCBI offers two options for species-specific protein databases: all available sequences or Reference Sequences (RefSeq). The former are not appropriate choices for proteomics use due to size and redundancy; only the RefSeq databases should be used for database searching. IPI offered only a single database choice for each available organism. Ensembl offers two FASTA files for each of its supported species. The “all” databases contain validated coding regions, whereas the “abinitio” files have predicted coding regions and should be avoided.

2.2. Protein Database Processing

Python scripts were used to extract redundant and non-redundant protein counts for each protein sequence database. Some scripts were available from www.ProteomicAnalysisWorkbench.com and others were written for this project. Redundant proteins were counted by comparing full-length protein sequences for string identities. Non-redundant protein databases were obtained by removing duplicated protein sequences (if any). *In silico* tryptic digestions were computed using regular expressions to cleaved proteins at arginine (R) and lysine (K) residues except when a proline (P) followed either one of the residues. The regular expression used (`r".(?:(?<![KR](?!P)).)*"`) correctly includes the protein N-terminal and C-terminal peptides. Additional processing of the tryptic peptides allowed for missed cleavages (up to a maximum of two) and required tryptic peptides to be at least seven amino acids in length. Unique and shared peptide counts were computed by tallying the number of protein sequences associated with each tryptic peptide.

2.3. Biological Proteomic Datasets

We used publicly available datasets of human peripheral plasma samples (Keshishian et al. 2015), TMT-labeled mouse c-Kit expressing cultured cells (Huan et al. 2015), and yeast BY4741 whole cell lysates (Hebert et al. 2014). Sample and mass spectrometry details can be found in the cited publications. The human sample RAW data (roughly 250 GB) was downloaded from the link in the publication and converted to MS2 format (McDonald et al. 2004) using MSConvert from the Proteowizard toolkit (Kessner et al. 2008) and in-house Python scripts that are part of the PAW pipeline (Wilmarth et al. 2009) used in the OHSU Proteomics Core. The dataset produced a total of 4,338,818 MS/MS spectra, which were centroided before database searches. The mouse data was used by permission of the authors and consisted of 272,221 MS/MS spectra. Instrument files for the yeast samples were downloaded via links in the publication and converted in a similar fashion to the human samples. The four LC runs performed under similar conditions were used for the analysis and represented a total of 318,069 MS/MS spectra.

2.4. Standardized Data Analysis

The PAW pipeline (Wilmarth et al. 2009) was used to provide a standard processing framework for the biological samples. This pipeline performs post processing of database search results and employs a best practices philosophy toward data analysis. An open source version of SEQUEST (Eng et al. 1994), called Comet (Eng et al. 2013) was used to perform the database searching. The protein databases listed in Table 2.1 were used in the searches. A wide parent ion monoisotopic mass tolerance of 1.25 Da was used with accurate mass post filtering to increase sensitivity (Hsieh et al. 2009). Fragment ion monoisotopic mass tolerances were 1.0005 Da. The human samples were iTRAQ (Ross et al. 2004) labeled and the mouse samples were TMT (Thompson et al. 2003) labeled. The isobaric reagent masses were specified as static (fixed) modifications as was alkylation of cysteine. Oxidized methionine was specified as a variable modification. Tryptic cleavage was specified with a maximum of two missed cleavages. Scoring used y- and b-ions in addition to neutral loss ions.

The target/decoy strategy (Elias and Gygi 2007) was used to estimate PSM error rates and set score filtering thresholds. Reversed protein sequences and common laboratory contaminants were added to each protein database before searches were performed. Optimum separation between correct and incorrect PSM score distributions was achieved by using Comet score transformations similar to those used in Keller et al. 2002. False discovery rate analysis was performed independently across subclasses of peptides along the lines of Ma et al. 2009 to ensure accurate overall FDR control. Parsimonious protein inference was performed using Python set operations and in-house scripts. An additional extended parsimony clustering module is described below.

2.5. Results Processing

There were detailed protein identification and peptide identification reports produced by the PAW pipeline after basic parsimony analysis and also after extended parsimony clustering. Scripts were written to parse the protein reports, remove any matches to common contaminants or decoys, count the total number of proteins in the report (redundant protein count), and count the net number of proteins/protein groups (non-redundant protein count).

The companion peptide reports were parsed and, in conjunction with the protein reports, the shared and unique status of peptides determined for three contexts: with respect to the original protein sequence database, with respect to the basic parsimonious protein list, and with respect to the final protein list after extended parsimony clustering (described below). Total numbers of identified peptides, total numbers of unique peptides, and total numbers of shared peptides were tabulated.

2.6. PAW Clustering Algorithm

The Proteomic Analysis Workflow (PAW) clustering algorithm was originally developed in 2010 when it was recognized that many common housekeeping genes did not have reliable unique spectral counts assigned to them after the basic parsimony analysis. Proteins in those well-known families such as actins, tubulins, keratins, heat-shock proteins, histones, etc. often had very large numbers of common peptides shared between family members with few formally defined unique peptides. In many cases, individual family members, while having sufficient evidence for identification, did not have sufficient unique peptide evidence for reliable spectral counting (counting the total number of fragmentation spectra that map to peptides of a specific protein) quantification (Liu et al. 2004; Lundgren et al. 2010). However, the family as a whole could be quantified if the family members were clustered together, and shared and unique peptides were subsequently redefined.

After the basic protein inference process, the Python algorithm performs pairwise comparisons of peptide sets associated with proteins that share peptides to decide if they should be combined into a single cluster. Pairs of proteins are clustered together based on shared and unique peptide spectral counts by one of the three tests: if the peptide sets are pseudo-redundant, if one set is a pseudo-subset of the other, or when overall relative shared peptide evidence dominates.

The test to cluster proteins as pseudo-redundant is to identify those proteins that have just enough unique peptide evidence to escape the peptide set equality test employed by the basic parsimony logic during the protein inference process. Two proteins are clustered as pseudo-redundant if they have a very low unique peptide spectral count (maximum of 2.0) but have a significant shared peptide spectral count (at least 10 times unique peptide spectral count of either protein). The idea behind this test is that if the marginal unique peptide evidence hadn't existed, then both the proteins would have been identified as truly redundant during the protein inference process.

A protein is clustered as a pseudo-subset of another protein if that protein has low unique peptide evidence (maximum of 2.0) and both proteins have considerable shared peptide evidence (at least 10 times the unique spectral counts of the tested protein). This test, like the pseudo-redundant test, is designed to catch cases where subsets evade the parsimony logic used during the protein inference process by having some, but not a sufficient amount, of unique evidence.

Finally, the third test is to cluster proteins based on total shared peptide evidence. In this test, a pair of proteins are clustered together if one protein has low mean unique spectral count per experiment (maximum of 2.5) and a mean shared spectral count per experiment greater than a threshold value (shared threshold of 20.0) or if the protein's shared-to-unique total spectral count ratio is greater than a threshold value (shared to unique threshold of 40.0). This test is more general than the previous two, which address situations where random incorrect PSMs may interfere with the basic parsimony analysis, and looks for cases where the shared evidence overwhelms the unique evidence. Whenever a pair of proteins form a cluster, they are combined into a single entry and the unique and shared peptide counts are re-computed. The clustering iteration continues until a stable number of clusters are generated.

2.7. Scaffold-like Clustering Algorithm

Scaffold (Searle 2010) (Proteome Software, Inc., Portland, OR) is a commercial package that provides users with an all-in-one application implementing PeptideProphet (Keller et al. 2002) and ProteinProphet (Nesvizhskii et al. 2003) algorithms to identify the most likely proteins present in proteomics experiments. The Scaffold protein clustering algorithm is an extension to its protein inference process. The clustering algorithm assembles proteins into clusters based on shared peptide evidence. Similar to the PAW algorithm, Scaffold clustering performs pairwise comparisons of peptide sets and decides whether to cluster or not based on shared peptide evidence.

The Scaffold algorithm clusters a pair of proteins if two criteria are satisfied: Firstly, the sum of the probabilities of the shared peptides of both the proteins must be at least 95%. The probabilities of the peptides, generated using PeptideProphet, are Bayesian estimates of the probable confidence of identified PSMs from the database search. Secondly, the proteins must share at least 50% of their peptide evidence. The probabilities of shared peptides are summed and compared with the total summed probability of all the peptides for each protein. If the sum of the probability of the shared peptides is greater than or equal to half the sum of the total peptide probability of all peptides for either protein, then the two proteins are clustered together. Every additional protein is then iteratively added to an existing cluster if that protein passes the above-mentioned criteria with any of the proteins present within a cluster.

In order to compare the two clustering algorithms, it is important to have both start with the same input data. There are many steps in proteomics pipelines and it is generally difficult to compare them (Yates et al. 2012) without careful controls. With this in mind, we altered the Scaffold algorithm from its described guidelines (https://proteome-software.wikispaces.com/file/view/scaffold_protein_grouping_clustering.pdf) to fit into the existing PAW pipeline to make both the algorithms comparable. The PAW processing does not assign Bayesian PSM identification probabilities; however, confidently identified peptides (FDR < 0.01) generally have greater than 0.95 probabilities in Scaffold. We assigned the confident peptides from the PAW processing to have probabilities of 1.0 in the calculations.

2.8. Internal Cluster Evaluation

Global pairwise alignments of proteins within clusters generated by PAW or Scaffold-like algorithms were performed and the sequence similarity score with BLOSUM62 substitution matrix was computed using Biostrings (v2.40.2) (Pages et al. 2008) software package from R Bioconductor (Release 3.3). Pairwise dissimilarity scores were computed from the sequence similarity score of each pairwise alignment. Python scripts were generated to compute the mean silhouette score using the dissimilarity score as the distance metric.

2.9. Functional Relatedness of Clusters

The Gene Ontology (GO) Consortium has produced a structured, well-defined, controlled vocabulary (ontologies) for describing the roles of genes and gene products in any organism (Ashburner et al. 2000; Gene Ontology Consortium 2004). GO ontologies are represented as a hierarchical graph, where nodes in the higher levels refer to GO terms that have a broader meaning such as ‘signal transduction’ or ‘enzyme’ and nodes in the lower levels refer to GO terms that are more specific such as ‘pyrimidine metabolism’ or ‘adenylate cyclase’. Due to the vagueness of the term “function” when applied to genes or proteins (descriptions can range from biological activities to cellular structures), the GO Consortium has developed three different ontology structures: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

Molecular Function is defined as biochemical activities of genes or gene products. Ontologies of MF describe activities that perform the biochemical actions such as catalytic or binding activities. An example of a broad level MF ontology is ‘kinase activity’ or ‘transporter’ while a lower levels includes activities such as ‘Toll receptor ligand’ or ‘6-phosphofructose kinase activity’. Biological Processes refer to the biological objectives accomplished by gene or gene products. A broader example of BP ontologies include ‘apoptosis’ or ‘cell growth and maintenance’ while a lower level BP ontologies include ‘cAMP synthesis’ or ‘apoptotic chromosome condensation’. Cellular Component refers to the place in the cell where the gene or gene product is active. Higher level nodes of CC include terms such as ‘ribosome’ or ‘proteasome’, while lower level nodes include more specific regions such as ‘ubiquitin ligase complex’. GO ontologies are often used for functional or biological enrichment (also known as gene enrichment analyses), where the analyses finds which GO terms are over- or under-represented in a given gene or protein set compared to a background set.

Functional relatedness of the biological clusters generated from PAW or Scaffold-like algorithms were tested using gene enrichment analysis. Gene enrichment analysis was performed using the GOATOOLS (v0.6.5) Python package (https://github.com/tanghaibao/goatools). GOATOOLS requires GO ontologies and their associations with a gene for a given species. GO ontologies for human, mouse, and yeast were downloaded (http://geneontology.org/ontology/go-basic.obo). GO annotations were retrieved from NCBI gene2GO ftp link (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz). NCBI gene2GO annotations list GO terms assigned to genes based on their Entrez Gene IDs. Protein accessions were converted to Entrez Gene IDs using biomaRT (v2.28.0) software package (Durinck et al. 2005; Durinck et al. 2009) from R Bioconductor (Release 3.3). Uncorrected *p-values* are computed for each GO term corresponding to BP and MF ontologies within the protein clusters using the Fisher exact test. These *p-*

values are further corrected for multiple testing using the Benjamini-Hochberg FDR method (Benjamini and Hochberg 1995) with a restriction of 0.05. A cluster is considered to have an enriched GO term if that term is associated with all the members of the cluster and has a corrected *p-value* of less than 0.05.

2.10. Statistical Testing

Two-sided paired t-tests were performed to test for significant effects of clustering on numbers of protein and peptide identifications. Other statistical tests such as two-sided F-test for equal variance was also performed to test if clustering has any effect in the variability of the numbers of protein and peptide identifications across protein sequence databases. The null hypotheses were rejected when the significance was less than 0.05.

Results and Discussion

3.1. Protein Sequence Database Analyses

Protein sequences were obtained from four frequently used protein sequence database sources (UniProt, NCBI, IPI, and Ensembl) for both higher eukaryotic (human and mouse) and lower eukaryotic organisms (yeast) (see Table 2.1). The Uniprot database consists of two sections: manually reviewed Swiss-Prot and automatically annotated TrEMBL. The relationship between Swiss-Prot and TrEMBL is complicated and is different for different species. For some organisms, such as human and mouse, Swiss-Prot databases are reasonably complete and can be used in proteomics studies. For most organisms, Swiss-Prot can be incomplete and should be augmented with TrEMBL sequences. Since the retirement of IPI databases in 2011 (Griss et al. 2011), UniProt has offered more filtered combinations of Swiss-Prot plus TrEMBL databases known as Reference Proteomes. Lower quality TrEMBL sequences and protein fragment sequence are filtered out. Protein isoforms are represented differently in Swiss-Prot and TrEMBL. In TrEMBL, isoforms are represented explicitly as different protein sequences. Isoforms are annotated as sequence differences to a canonical sequence selected for each gene in Swiss-Prot. FASTA format protein databases for reviewed sequences (Swiss-Prot) can consist of just canonical sequences or a combination of canonical sequences and their isoforms. For species like human and mouse with more complete Swiss-Prot databases, large numbers of protein isoforms are no longer present in TrEMBL and only available in FASTA files if the correct download options are selected. Because of the complicated structure of UniProt, six database variants were used for human and mouse, and four for yeast. For Ensembl, NCBI, and IPI we used species-specific reference proteomes as sequence databases in our analyses.

Redundant and Non-Redundant Protein Analysis

The databases were processed to understand the differences between these sources. The total sequence counts for each database were computed for both higher and lower eukaryotic organisms and are shown in Table 3.1. The databases were also checked for duplicated, identical sequences. The only databases that were truly non-redundant (no duplicated protein sequences) were IPI for human and mouse. According to documentation from UniProt, TrEMBL database should include only a single record for an identical, full-length protein sequence from an organism, and Swiss-Prot should contain only one record per gene in one species. Swiss-Prot databases had small numbers of duplicates, the human and yeast Reference Proteomes had small numbers of duplicates, the mouse Reference Proteome had about 4% duplicated sequences, and that full TrEMBL databases result in larger numbers of duplicates for human and mouse but not yeast. NCBI RefSeq and Ensembl for human and mouse had large numbers of duplicate protein sequences perhaps reflecting their genomic focus where a variety of genomic and transcriptomic events can result in identical protein sequences. Yeast had a low proportion (about 1%) of duplicate copies of protein sequences for all databases.

There was a greater variability in non-redundant protein counts across all databases for higher eukaryotic organisms compared to lower eukaryotic organisms as shown in Figure 3.1. For example, non-redundant protein counts for humans ranged from 20,146 to 132,398 proteins (**mean:** 84,521, **SD:** 41,350, **CV:** 48.9%) while for yeast, the range was only between 5,844 to 6,665 proteins (**mean:** 6,510, **SD:** 327, **CV:** 5.0%). Larger human protein databases such as Swiss-Prot + TrEMBL + isoforms have over twice as many proteins as the Uniprot Reference Proteome, while smaller protein databases such as Swiss-Prot are only about 30% as big as the Uniprot Reference Proteome. However, little variation was observed in yeast, suggesting that the variations could be due to the inherent genomic complexity between higher eukaryotic organism (humans and mice) and lower eukaryotic organisms (as represented by yeast).

Redundant and Non-Redundant Tryptic Peptide Analysis

It is also important to consider protein databases from a peptide-centric view given the wide spread use of shotgun proteomics. *In silico* tryptic digestions were performed on the non-redundant protein databases (after removing duplicate proteins) for higher and lower eukaryotic organisms. Simulated tryptic digestions cleaved all protein sequences at arginine (R) and lysine (K) residues if they were not followed by proline (P), and allowed up to two missed cleavages. Tryptic peptides with fewer than seven amino acids were ignored. The total numbers of distinct peptide sequences (the digested peptide set size) were tabulated. Peptides originating from a single protein sequence were classified as unique and peptides that came from multiple proteins (leucine and isoleucine were considered indistinguishable) were classified as shared. The data from the analyses are presented in Table 3.2.

The sizes of the *in silico* digestions for human and mouse were considerably larger than yeast and varied more database to database. For example, the variability in digestion sizes across human databases (**CV:** 9.5%) was greater than for yeast databases (**CV:** 1.1%). In contrast to the considerable variation in non-redundant protein counts across human databases, the peptide counts are more similar, as can be seen in Figure 3.2. For instance, the Swiss-Prot + TrEMBL + isoforms database had 2.2 times as many non-redundant proteins as the Uniprot Reference Proteome, but had only 18% more distinct peptides.

It is also apparent from Figure 3.2 that higher eukaryotic organisms have greater proportions of shared peptides compared to lower eukaryotic organisms. Human protein databases revealed that, on average, 53% of all peptides originated from multiple proteins, while shared peptides were only 1.9%, on average, across yeast protein databases. The variability in shared peptide proportions across human sequence databases was large (**CV:** 37.9%), although smaller than for protein sequence counts. The Swiss-Prot databases for human and mouse illustrate quite clearly the effects of manual curation where related proteins are grouped together and replaced with single canonical sequences. The 2-3% shared peptide levels in Swiss-Prot databases (Figure 3.2) are dramatically smaller than that of any of the other human or mouse databases. Yeast databases exhibited less variation in shared peptide proportions (**CV:** 13.8%), probably an indication of the reduced genomic complexity of this lower eukaryotic organism.

3.2 Analyses of Biological Samples

Biological samples seldom contain all the proteins present within sequence databases. It is possible that proteins identifiable in biological samples may have protein characteristics that differ from those computed from all proteins present in protein sequence databases. We selected three representative biological studies with datasets available for analysis: human peripheral plasma samples, TMT-labeled mouse c-Kit expressing cultured cells, and yeast BY4741 whole cell lysates. The studies were performed on state-of-the-art instrumentation and have large datasets capable of challenging processing algorithms. Two of the experiments (human and mouse) used isobaric labeling for quantification and represent the two most commonly studied higher eukaryotic organisms in biomedical research. Yeast is a widely used system in proteomics for method development and a good choice for a representative lower eukaryote lacking the genomic complexity of human and mouse.

Comparing alternative proteomic data processing results turns out to be surprisingly difficult (Yates et al. 2012) due to the large number of processing steps, each of which often have several adjustable parameters. In light of these realities, a single processing pipeline, namely, the PAW processing (Wilmarth et al. 2009) was used to provide standardized data analyses for the three biological samples. The same processing steps with controlled parameter choices (details are given in Section 2.4) could be applied to all analyses. Separate PAW analyses were done using each of the protein databases listed in Table 2.1.

The PAW pipeline produces parsimonious protein reports that detail all identified proteins and protein groups meeting basic protein identification criteria. The protein identification criteria was the widely-used two peptide rule (Gupta and Pevzner 2009) where proteins were required to have two distinct, confidently identified peptides present for each reported protein in each biological sample in an experiment. Note that any proteins associated with peptide subsets that are removed during protein inference are not included in the PAW reports. The human and yeast experiments each had four biological samples per experiment; the mouse experiment was a single sample that was a mixture of eight TMT-labeled biological samples. Due to the peptide centric approach of shotgun proteomics, the protein inference process results in redundant (indistinguishable) and non-redundant (distinguishable) proteins. Note that the definitions of redundant and non-redundant in this (historical) proteomics context is not the same as redundant and non-redundant proteins in terms of protein sequence databases used in Section 3.1. Redundant protein groups, in this context, are groups of proteins that have been identified from the same set of peptides, whereas non-redundant proteins are identified from a distinct (not necessarily exclusively unique) set of peptides.

Protein Characteristics of Biological Samples

For each species-specific protein sequence database, the PAW protein inference pipeline generated a final list of proteins (redundant groups and non-redundant proteins) that have been identified from peptide evidences from the biological samples. Prior to computing redundant and non-redundant protein counts, all redundant protein groups that contained either contaminant or decoy proteins and all contaminant or decoy non-redundant proteins were discarded. The PAW result file analyses are tabulated in Table 3.3.

Not surprisingly, the numbers of identified proteins differed between the three samples. On average, there were 6,249 total proteins identified in human plasma samples across all databases compared to an average of only 3,815 total proteins identified in yeast whole cell lysates. More importantly, there was also higher variability in total proteins counts in real samples of higher eukaryotic organisms across sequence databases than yeast whole cell lysates (**CV Human total:** 28.5%, **CV Mouse total:** 24.8%, **CV Yeast total:** 0.2%). From inspection of Table 3.3, it is clear that protein database choice is not too critical for yeast. This organism has been thoroughly studied and the major database sources actively communicate with each other. Thus, it is likely that the yeast protein databases are very similar from the various sources.

The protein databases for human and mouse have greater size differences than those for yeast. The total numbers of identified proteins (redundant counts) are correlated with database size, as shown in Figure 3.3(A). The non-redundant proteins also have dependence upon the protein database size, but to a much lower degree (Figure 3.3(B)), as judged by the differences in slopes of the trend lines. The situation for mouse is quite surprising where the non-redundant protein counts are nearly constant despite large differences in redundant counts. This could indicate that alternative protein forms for mouse are more similar after tryptic digestion than for human and, therefore, more difficult to distinguish in shotgun proteomic studies.

Along these lines, the canonical Swiss-Prot databases for human and mouse resulted in dramatically fewer redundant protein groups than the other protein databases suggesting that the canonical sequences are, for the most part, distinct. It is also notable that the gain in non-redundant protein identifications is very minor when comparing Swiss-Prot to Swiss-Prot + Isoforms (the bottom two rows) in Table 3.3. There are 22,000 additional isoforms being searched for human and 8,000 for mouse. The human plasma dataset is, literally, enormous with over 4 million MS/MS spectra, yet the data supports barely more than 100 protein isoforms out of over 3000 identified proteins. This suggests that identifying distinguishing peptides from protein isoforms in typical shotgun proteomics experiments is extremely difficult.

The PAW pipeline is designed to be very transparent and its detailed log files allow tracing of all steps in the protein inference processing. It is not well understood, nor appreciated, just how challenging large datasets and large protein databases are on proteomics data processing. Table 3.4 details protein inference for the human samples across the nine protein databases. The PSMs are strictly filtered at a 1%

false discovery rate prior to protein inference in all cases. For this large dataset with a very wide dynamic range sample (plasma), the numbers are striking. More than half of the initial proteins are detected by single peptides. PSM errors directly produce protein identification errors for this protein class, and are the major reason that they are routinely excluded from results lists. Indistinguishable peptides sets account for nearly another factor of two reductions in protein count, except for the more curated databases without isoforms. There is a dramatic reduction in protein numbers after subset removal for all databases except Swiss-Prot. This is a major reason that the Paris Guidelines for Proteomics Results (Bradshaw et al. 2006) were created. It is all too easy to inflate the number of identified proteins with proteins lacking any true experimental evidence.

3.3. Extended Parsimony Clustering

The number of non-redundant protein/protein groups reported by the PAW pipeline should represent a parsimonious list of identified proteins. For each sample, the same data was used and the dependence of the protein identifications on the choice of protein database was explored in the preceding section. Given that these three organisms are so well studied, it is reasonable to assume that the majority of proteins present in the sample would have representation in the protein databases. Thus, the truly parsimonious number of identified proteins should be (more or less) dependent on the sample only and be relatively independent of protein database choice. The difference between the largest non-redundant protein count and the smallest was 1560 for human, 358 for mouse, and 5 for yeast. This suggests that the larger protein databases for human or mouse, their more complicated gene structures, or the all-or-nothing equality testing used in basic parsimony logic might play roles in the variable numbers of protein identifications.

The dataset size for the human samples is much larger than for mouse or yeast. All datasets are filtered at the same relative error content, namely, a 1% PSM false discovery rate. The average number of MS/MS scans associated with decoy proteins for human was roughly 8,800 per analysis in contrast to an average of just 900 for mouse. Dataset size is another important factor to consider.

The variable numbers of identified proteins for the same samples, particularly human, suggests that the factors mentioned above may be "breaking" the basic parsimony logic consistent with recent work by Koskinen et al. 2011 where hierarchical clustering was used to provide an additional round of protein grouping beyond the basic parsimony logic. Use of traditional geometric clustering algorithms requires definition of an appropriate distance metric. The Mascot (Perkins et al. 1999) ion score for unique peptides was used in Koskinen et al. 2011. This has some limitations and more work in this area is needed to determine the best distance metrics. A popular commercial proteomics analysis package (Scaffold, Proteome Software, Inc., Portland, OR) developed an alternative algorithm based on comparisons of identified peptide sets that is described in this document: (https://proteome-software.wikispaces.com/file/view/scaffold_protein_grouping_clustering.pdf). That algorithm is similar to an independently developed algorithm available in the PAW pipeline. These later two algorithms, which make use of peptide set comparisons to perform extended parsimony clustering, could be implemented within the same processing framework and were compared to basic parsimony results and to each other. The non-redundant protein identification numbers before and after running the two clustering algorithms are listed in Table 3.5.

PAW Extended Parsimony Clustering

The PAW clustering algorithm parsed the protein inference reports generated from the PAW pipeline and clustered proteins based on relative shared and unique peptide evidence. We ran the PAW clustering algorithm on the protein inference reports for each sequence database and Table 3.6 shows the cluster summary statistics from human plasma samples and yeast whole cell lysates. We have observed that the PAW clustering algorithm generated significantly more clusters on human samples than yeast samples (**Two-sided *p-value*** = 0.0012). Human plasma samples, on average, generated 231 clusters across multiple protein databases, while yeast whole cell lysates, on average, generated only 14.5 clusters. As observed with non-redundant protein counts, there appeared to be a greater variability in protein clusters in human samples (**CV**: 55.2%) compared to yeast samples (**CV:** 3.7%).

There appeared to be a strong association between non-redundant protein counts detected in human plasma samples prior to PAW clustering and the cluster counts after implementing PAW clustering (**R^2 =** 0.94, **F-test *p-value*** < 0.001) (Figure 3.4). Human plasma sample protein lists from larger protein databases generated more clusters than those generated from smaller databases. Figure 3.5 shows that this positive association compensates for the effects of sequence database sizes on final non-redundant protein counts.

There is a reduction in the variations in non-redundant protein counts in human plasma samples after PAW clustering: pre-clustering **CV:** 16.0%, post-clustering **CV:** 5.2%, and **two-sided F-test for equal variance *P-value*** = 0.002. PAW clustering had little effect on the numbers of identified proteins observed for yeast whole cell lysates. The largest effects of PAW clustering on human non-redundant protein counts were observed in the largest databases: Swiss-Prot + TrEMBL with and without isoforms.

Scaffold-like Extended Parsimony Clustering

We also implemented a Scaffold-like clustering algorithm (see Methods section 2.7) and post-processed all of the results files for the three biological samples searched against the different protein sequence databases. The trends observed with Scaffold-like clustering on protein and peptide characteristics were similar to those observed with PAW clustering. We observed significantly more clusters generated on human plasma samples compared to yeast whole cell lysates (**Two tailed paired t-test *P-value* <** 0.001). From the data presented in Table 3.6, Scaffold-like clustering, on average, generated 299 clusters in human plasma samples across human sequence databases, while only 61 clusters, on average, were generated in yeast whole cell lysates. We noticed a greater variability in the protein cluster counts for human samples compared to yeast samples. Scaffold-like clustering, similar to PAW clustering, significantly reduced the variability in the non-redundant protein counts for human samples, as shown in Figure 3.6. Post-clustering, the variation in non-redundant protein counts decreased from 16.0% to 0.9% (**Two-sided F-test for equal variance *P-value*** < 0.001), demonstrating the stabilizing effect of extended parsimony clustering on protein identification numbers.

Scaffold-like clustering was, in general, more aggressive than PAW clustering. At its default settings, Scaffold algorithm generated, on average, 34.5% more clusters with human plasma samples across all human sequence databases than the PAW clustering algorithm at its default settings. This could perhaps be due to the Scaffold-like algorithm’s more relaxed criteria for clustering. Due to this, the Scaffold-like clustering generated significantly more clusters than PAW clustering for both human and yeast samples (**Two sided paired t-test *P-value*** < 0.001), as can be seen in Figure 3.7. Not only does Scaffold-like clustering generate more clusters (singletons excluded), but also the mean cluster sizes are greater for human samples than for PAW clusters (**Mean cluster size (PAW):** 2.6, **Mean cluster size (Scaffold-like):** 3.4).

Interestingly, the mean cluster sizes are lower for yeast samples with Scaffold-like clustering than for PAW clustering. Another interesting observation, shown in Figure 3.8, is that Scaffold-like cluster sizes varied more widely than cluster sizes from PAW clustering. For instance, we observed several clusters sizes ranging from two to over ten proteins with Scaffold-like clustering in human samples for the Uniprot reference proteome (canonical) database, but PAW clustering only generated cluster sizes of two, three, four, five, and eight. It is interesting to note that the Scaffold-like algorithm, at an 80% shared peptide threshold, appeared to have generated a similar number of clusters for human plasma samples across all databases as PAW clustering algorithm at its default settings (Figure 3.9).

Cluster Evaluation

A true evaluation of the two clustering methods would be to compare their outputs to those from a truly accepted clustering algorithm. However, such a truth clustering set doesn't exist for clusters based on mass spectrometric information. Alternatively, the two cluster methods could be evaluated based on valid internal cluster metrics such as Silhouette score and external attributes of clustered proteins such as enrichment analysis of functional annotations.

Internal Evaluation of PAW and Scaffold-like Clustering

Internal evaluation of the clustering algorithms was performed by computing the mean silhouette scores of clusters generated for human and yeast samples from multiple species-specific protein sequence databases. The mean silhouette score computes a measure of how similar a protein is to its own cluster compared to its neighboring cluster (Rousseeuw 1987). It ranges from -1 to 1 with higher mean silhouette score suggesting that the proteins are more tightly grouped to their respective clusters. The dissimilarity scores from global pairwise alignments of all pairs of clustered proteins identified in PAW or Scaffold-like algorithms was chosen as the distance metric for computing silhouette score. The internal evaluation procedure is illustrated in Figure 3.10.

The mean silhouette scores were high for clusters generated by either algorithm for human or yeast samples and are shown in Figure 3.11. The mean silhouette score was 0.48 for the PAW algorithm for human samples across all sequence databases, while Scaffold-like clustering generated a mean silhouette score of 0.59 for the same samples across all databases. There appears to be a lower variability in mean silhouette scores for human samples with Scaffold-like clustering (**CV:** 10.9%) compared to that with PAW clustering (**CV:** 37.7%). Interestingly, the biggest discrepancy in mean silhouette scores between PAW and Scaffold-like clustering appeared to be in the extremes of database sizes such as Swiss-Prot + TrEMBL or Swiss-Prot (see Figure 3.11). Mean silhouette scores between PAW and Scaffold-like methods had a noticeable difference in yeast samples. The mean silhouette score was 0.44 with the PAW algorithm in yeast whole cell lysates, whereas the Scaffold-like algorithm generated a higher mean silhouette score of 0.75 across all yeast sequence databases.

External Evaluation of PAW and Scaffold-like Clustering

Another likely feature of a good clustering method would be that proteins grouped together in the same clusters would have similar biological functions. One way to check for this is to use gene enrichment analysis to test if the clusters generated from each algorithm shared any significant gene ontology (GO) terms for biological process (BP) and molecular function (MF) ontologies.

Entrez Gene IDs were determined for all proteins within a cluster generated from PAW or Scaffold-like algorithms for human and yeast samples (see Methods section 2.9). Using their respective Entrez Gene IDs, the GO terms for BP and MF ontologies within each cluster could be compiled by GOATOOLS. All the proteins generated with a species-specific sequence database in the protein inference report were considered as the background set for the Fisher Exact test. Uncorrected *p-values* were computed for each GO term in a cluster and further corrected for multiple testing using the Benjamini-Hochberg FDR method with a FDR restriction of 0.05. Clusters were considered eligible for gene enrichment analysis only if the cluster had at least two proteins with unique Entrez Gene IDs. A cluster could contain GO terms that are either enriched (protein cluster has significantly higher concentration of a GO term compared to the background) or depleted (protein cluster has a significantly lower concentration of a GO term compared to the background). A cluster was considered to have a significant GO term only if all the proteins of that cluster have an enrichment of that GO term (all the proteins of the cluster are annotated with that GO term and have a corrected *p-value* of less than 0.05).

Only a small proportion of clusters met the criteria for GO enrichment analysis despite a large number of clusters generated by the clustering algorithms. Of all the clusters generated for human samples listed in Table 3.7 across all sequence databases by the PAW algorithm, we computed that, on average, only 23.9% of them were eligible for gene enrichment analysis. A similar low eligible fraction (36.3%) was also observed for human protein clusters generated using the Scaffold-like algorithm (Table 3.7). However, a surprisingly large proportion of the eligible clusters had significantly enriched GO terms for BP or MF ontologies (Figure 3.12) for both PAW and Scaffold-like clustering. About 73.2% of the eligible clusters contained at least one significantly enriched GO term for BP or MF ontologies with PAW clustering, while 76.9% of the eligible clusters contained at least one significantly enriched GO term for BP or MF ontologies with Scaffold-like clustering. This suggests that the clustering algorithms produce clusters that have a biological relatedness; however, annotation limitations severely reduced the numbers of clusters eligible for the gene enrichment analysis.

3.4. Quantitative Information Content

The quantitative information content (QIC) of a proteomics experiment will be defined here as the fraction of total unique peptide PSMs out of the total number of confidently identified PSMs. However, protein inference and parsimony logic changes the context in which unique and shared peptides are defined. The extended parsimony clustering algorithms create a third context within which shared and unique peptides can be defined. As was mentioned in the introduction, shared peptides are ambiguous, and the most common treatment of shared peptides in quantitative proteomics is to discard them. Thus, QIC will depend on protein context and may depend on protein database choice. The PAW pipeline, including the extended parsimony clustering step, generated reports of all the peptides identified from mass spectrometry and could be used to compute the numbers of shared and unique peptides in these different contexts.

Peptide Properties in Biological Samples

Interestingly (and by chance), the human plasma samples and yeast whole cell lysates generated similar total mean numbers of identified peptides with a very low variance across their respective species-specific sequence databases (**Human peptides: mean:** 46,815, **CV:** 0.8%; **Yeast peptides: mean:** 46,492, **CV**: 0.004%). The peptide reports produced by the PAW pipeline, without employing the extended parsimony clustering step, list peptides with shared and unique status defined with respect to the parsimonious protein list. Human samples appeared to have a significantly greater percentage of shared peptides compared to those of yeast samples as can be seen in Figure 3.13. Human samples, on average, had about 13.5% of shared peptides compared to yeast samples, which, on average, only have about 2.8% shared peptides (**two sample t-test: *p-value*** < 0.001). Yeast whole cell lysates produced similar shared peptide proportions to those observed from *in silico* digestions of yeast sequence databases (mean shared peptide proportion from *in silico* digestions: 2.2%); however, *in silico* digestions of human sequence databases had a higher average of 54% shared peptides.

Peptide Properties due to Extended Parsimony Clustering

We have observed a significant decrease in shared peptide proportions upon implementing PAW clustering, particularly on human plasma samples. The mean shared peptide proportions across databases for human samples reduced from 13.5% pre-clustering to 4.3% post-clustering (**Two-sided paired t-test *p-value* <** 0.001). Also, as seen in Figure 3.14, PAW clustering has reduced the variability in shared peptide proportions across databases (**Two-sided F-test for equal variance** *p-value* < 0.001). Scaffold-like clustering appeared to have a similar effect as PAW clustering on shared peptide proportions. The effects of Scaffold-like clustering on shared peptide proportions appeared to be more drastic as the mean shared peptide proportion across databases reduced from 13.5% pre-clustering to a mere 1.1% post-clustering as can be seen in Figure 3.15 (**Two-sided paired t-test** *p-value* < 0.001). Similar to PAW clustering, Scaffold-like clustering significantly reduced the variability in shared peptide proportions across databases (**Two-sided F-test for equal variance** *p-value* < 0.001). However, no effects of extended parsimony clustering on peptide characteristics were observed for yeast whole cell lysates.

QIC in Changing Protein Contexts

All confidently identified PSM numbers and the numbers of unique PSMs in the three different contexts are summarized in Table 3.8 for all combinations of datasets and protein databases. PSMs unique to the protein database are those PSMs that map to a single protein entry in the protein sequence database. PSMs unique after basic parsimony are PSMs that map uniquely to a single distinguishable protein or indistinguishable protein group within the set of reported protein identifications. The final two columns are PSMs that are unique to distinguishable proteins, indistinguishable protein groups, or extended parsimony protein clusters from either the PAW clustering or the Scaffold-like clustering.

Several observations can be made from Table 3.8. The total PSM counts are relatively independent of protein database choice for each biological sample. That is not too surprising since the same starting data is used for each database search. All numbers for yeast are essentially independent of protein database choice again suggesting that yeast protein databases from different source may be very similar. For human and mouse, the Swiss-Prot and Swiss-Prot + isomers databases resulted in total confident PSM numbers that were a little smaller that for the other databases. This suggests that there may be some classes of proteins missing from Swiss-Prot that are present in the other databases. The numbers of PSMs that were unique to the protein database, as expected, varied considerably for human and mouse, with the larger protein databases having fewer unique PSMs. While these numbers are low in many cases, the associated PSMs have no ambiguity about what protein they represent.

Unique and shared PSM definitions with respect to the protein database are not how proteomics data are typically reported. Definitions with respect to the list of identified proteins are more common. There are significant increases in the number of PSMs that are unique to the basic parsimonious lists of identified proteins for human and mouse compared to the protein database context. Interestingly, the mouse numbers are very stable with respect to protein database choice, whereas the human sample still has considerable protein database dependence. Both extended parsimony clustering algorithms further reduce protein database effects, particularly for the human plasma sample. The effect of extended parsimony clustering was much smaller for mouse than human, perhaps due to the much smaller dataset for mouse, or the nature of the mouse sample. The stabilizing effects of the extended parsimony clustering can be seen in Figure 3.16 where the human QIC data are shown.

Conclusions

Proteomic studies are very important in modern biology research. The majority of studies use bottom-up or shotgun proteomics where peptides rather than proteins are detected using mass spectrometry. Sequences are assigned to the detected peptides using search engines and protein sequence databases. Search engines have received extensive study, but only minimal work has been published on the role of protein sequence databases. Protein databases for human, mouse, and yeast were obtained from the major sources of protein sequence information, and, along with recent available proteomic datasets for each species, are used to understand the role of protein database on protein identification and quantification methods.

The sources for protein sequence databases were UniProt, NCBI, Ensembl, and IPI (excluding yeast) for these eukaryotic organisms. Protein databases for other organisms may include these sources or other sources. NCBI, Ensembl, and IPI have single database choices. UniProt is considerably more complicated. There are manually curated sequences (Swiss-Prot), computer annotated sequences (TrEMBL), and optional sequences for manually annotated protein isoforms. For human and mouse, there were six choices for UniProt databases, and four for yeast. The databases used are listed in Table 2.1 and described in the Methods and Datasets section.

The analysis of these different databases started with counting the number of protein sequences and checking for repeated, identical (redundant) sequences. The number of sequences and the fraction of redundant sequences was quite variable for human and mouse, and less so for yeast. Database sources that are more focused on proteins (UniProt and IPI) had fewer redundant proteins. Sources more focused on genomics (NCBI and Ensembl) had more protein redundancy. For higher eukaryotes like human and mouse, single-nucleotide polymorphisms, RNA editing, alternate splicing, gene duplications, and post-translational modifications can yield similar protein products from genes. There are two choices: one protein associated with multiple genes, or one gene associated with multiple proteins. This choice differed depending on database source. All of the duplicate counts in Swiss-Prot belonged to different genes that yielded identical protein sequences. Ensembl incorporates genomic, transcriptomics and proteomic data with separate database accessions for each branch of the central dogma. The redundant proteins in Ensembl were due to the presence of either different genes that produced a similar protein sequence or due to a gene having different transcript IDs or chromosomal locations that yielded the same protein sequence. Redundant protein sequences were removed from the protein databases so that subsequent analyses would not be biased.

Larger databases such as Swiss-Prot + TrEMBL had several times as many non-redundant proteins as SwissProt (canonical) for human and mouse; however, after *in silico* tryptic digestions of these non-redundant proteins, the largest protein database (Swiss-Prot + TrEMBL + isoforms) is only 33.6% larger in total peptide content than the smallest protein database (Swiss-Prot canonical). Given the peptide-centric nature of shotgun proteomics, it is important to also consider protein sequence databases from a peptide-centric point of view. There was great variability in the shared peptide proportions across human sequence databases in sharp contrast to the yeast databases. This could reflect the genomic complexity differences between higher and lower eukaryotic organisms, where additional post-transcriptional and post-translational processes can yield similar protein products in higher eukaryotic organisms. Interestingly, the Swiss-Prot human database had only 3.1% shared peptide proportion, while the Swiss-Prot + isoforms had 53.9% shared peptide proportions. A likely explanation of the higher shared peptide proportions in higher eukaryotic sequence databases could be due to the presence of large numbers of protein isoforms.

Proteomics datasets from three recent publications (Hebert et al. 2014; Huan et al. 2015; Keshishian et al. 2015) were used in searches against the different protein databases to see how they influenced protein inference. All datasets were from current mass spectrometry platforms and large enough to expose any weaknesses in standard proteomics data analyses. Standardized, best practices data processing used decoy databases (Elias and Gygi 2007), the Comet search engine (Eng et al. 2013), and the PAW pipeline (Wilmarth et al. 2009). The total protein counts after basic parsimony logic varied greatly across databases for human samples while remaining relatively stable across databases for yeast samples. The trends observed in the shared peptide proportions in these samples were very similar to those observed from *in silico* digestions of protein sequence databases.

Basic parsimony analysis was incapable of generating consistent protein identification numbers independent of protein database choice for the given samples, particularly the human plasma samples. It was possible that basic parsimony logic with simple equality-based testing could be failing when dataset sizes are too large and/or the protein databases have too much peptide degeneracy. Two available clustering algorithms that are basically extensions of parsimony logic were tested. The extended parsimony algorithms were applied after the protein inference step (that included basic parsimony logic) to cluster proteins with significant shared peptide evidence and relatively small unique peptide evidence. Basic parsimony principles are routinely used to report the minimum set of proteins that account for all the observable peptides to meet publication guidelines (Bradshaw et al. 2006). The clustering algorithms are extensions of the parsimony principles in that they cluster largely homologous proteins such as immunoglobins, MHC proteins, or housekeeping gene products (actins, tubulins, etc.) that are common in many samples.

The PAW clustering algorithm has three steps to test if there is insufficient unique peptide evidence to support distinguishing proteins. We coded an independent version of the Scaffold clustering algorithm (developed by Proteome Software, Inc., Portland OR) and compared it to PAW clustering. The Scaffold-like algorithm tests whether proteins have enough shared peptide evidence to be clustered together. Both algorithms make use of only experimentally measured information (peptide sequences, peptide scores, and peptide counts) in their tests and are computationally efficient.

The two clustering algorithms were evaluated for outcome quality by computing mean silhouette score of the clusters as an internal metric (using whole protein sequence alignments to derive distance measures), and by performing gene enrichment analysis to see that there is biological (function) relatedness among protein cluster members. Mean silhouette scores for each protein database were higher for both human and yeast samples with the Scaffold-like clustering algorithm, suggesting that the clusters from the Scaffold-like processing were more tightly packed than those from the PAW algorithm. The disparity in the mean silhouette scores was greatest in larger databases such as Swiss-Prot + TrEMBL for human samples. An explanation could be that the PAW algorithm generated multiple smaller clusters from similar protein classes (for instance MHC class I proteins) when they had some differences in their peptide sets. This would create situations where proteins within a cluster could match closely with both its own cluster and also a neighboring cluster. The Scaffold-like algorithm, on the other hand, is more likely to produce a single cluster consisting of all the proteins that belonged to the same class. In these cases, the neighboring clusters would belong to entirely different classes of proteins so there would be a lower possibility that a protein from a Scaffold-like cluster would be a better fit in its neighboring cluster.

Gene enrichment analysis was used to test if the clusters generated by either PAW or Scaffold-like algorithms had similar biological or molecular functions. Unfortunately, the gene enrichment analysis was limited because the majority of proteins within a cluster either belonged to either the same Entrez Gene ID (multiple proteins are translated from same genes) or did not have an Entrez Gene ID (no information). For example, the PAW clustering of human plasma samples from the Swiss-Prot + TrEMBL + isoforms database had only 19.5% (321 of 1646) of the proteins within a cluster (of size 2 or more) that had an Entrez Gene ID. There were 16% (51 of 321) of those proteins that had duplicate Entrez Gene IDs. The gene enrichment analysis could only be done on those clusters that had proteins yielding at least two unique Entrez Gene IDs, and a significant proportion of clusters had to be discarded (see Table 3.10). Despite the smaller number of testable proteins, most clusters did indeed have significant gene enrichment.

Shotgun proteomics experiments are widely used for quantitative studies. When protein expression is the goal, it is important to know unambiguously which protein that each peptide maps to. Peptides that can arise from multiple proteins potentially have expression measures that are a mixture of the respective protein expression levels and are difficult to interpret. Typically only unique peptides are used for quantification in shotgun studies. This raises the question of unique in what context. The answer really depends on the experimental goals. We studied here three different contexts within which unique peptide can be defined. The unique to the protein database context has the greatest quantitative resolution (the largest number of different protein forms that can be probed), but a reduced sensitivity. The loss of sensitivity depends quite strongly on the nature of the protein database and is worse for higher eukaryotic organisms with more complicated gene structure.

Another protein context is unique to the list of reported proteins (typically produced using basic parsimony logic). This is probably the most commonly used context in shotgun proteomics. It is important to realize that many peptides called unique in this context are not unique with respect to the protein database. Although this context results in many more usable PSM for quantitation (a higher QIC), as can be seen in Table 3.8, it comes at the cost of reduced protein resolution. For some samples and some choices of protein databases, extended parsimony clustering can further improve QIC (sometimes significantly, see Figure

3.16). Once again, this comes at the price of reduced protein resolution and the competing factors of a high QIC and high protein resolution have to be weighed against each other.

Finally, the obvious question of what is the best protein database to use needs to be addressed. The answer is easy for yeast since the protein database choice really did not matter much by any metric. For human and mouse, the answer depends on several factors. A common misconception might be that larger protein databases are better as they may be more complete and possibly yield more protein identifications. Since shotgun proteomics is peptide-centric, from analyses of *in silico* tryptic digests, both smaller and larger protein databases shared a significant proportion of peptides, which was a similar finding to what was also observed by Deutsch et al. 2016. For example, 96.5% (45,689) of the peptides detected in human plasma samples using the Swiss-Prot + TrEMBL + isoforms database were also present with Swiss-Prot canonical database. This suggests that only approximately 3.5% of the peptides were found exclusively in the larger Swiss-Prot + TrEMBL + isoforms database. Figures 3.5 and 3.6 demonstrate that non-redundant protein identification numbers, particularly after extended parsimony clustering, are pretty constant. This suggests that the experimental information content in most bottom-up studies may not yield larger protein identification numbers from larger databases. The sequence coverage for the majority of proteins detected in shotgun proteomics experiments is low and the chance to detect peptides that might distinguish protein variants is very small. Table 3.4 illustrates just how demanding large protein databases can be on the protein inference algorithms. The risk of using these large databases does not seem in line with the potential gains for most studies.

Proteomics experiments are very diverse and few generalizations are possible. There can be many experimental goals, even for the same experiment. There is no rule that only one protein database has to be used, or one single analysis has to be done for proteomics experiments. Different experimental goals may need use of more than one protein database. Ensembl databases have advantages when transcriptomics studies are being done in parallel with proteomics studies. Their larger sizes and protein redundancy; however, require proper protein inference and parsimony analysis steps. IPI database are out of date so should not be used; the recommended replacement databases are from UniProt. For human and mouse, Swiss-Prot databases are very good choices in most cases. For many (maybe most) other organisms, Swiss-

Prot sequences are seldom complete enough to use by themselves and have to be augmented with TrEMBL entries. It is pretty obvious from most of the data presented here that Swiss-Prot plus TrEMBL entries are poorer choices. Reference proteomes were introduced in late 2011 and are the better choices because of their filtered TrEMBL content. The way in which UniProt deals with protein isoforms is an important issue. For less well annotated organisms, isoform are likely present in TrEMBL. For well-annotated organisms like human, mouse, and yeast, isoforms are no longer present in TrEMBL and have to be (optionally) included in protein databases if they are of biological interest.

The expression levels of proteins are often of much greater biological importance than the longest list of identified proteins. For quantitative studies, the low peptide degeneracy of the canonical Swiss-Prot databases for human and mouse offer many advantages. They are more complete than might be guessed from their protein sequence counts. Most tryptic peptides have one-to-one relationships with the protein database entries and protein inference is greatly simplified. The proteins have rich annotations (protein functions, biological pathways, cross-references to other databases) that can greatly facilitate biological interpretations of results.

References

Aebersold, R. and Mann, M., 2003. Mass spectrometry-based proteomics. Nature, 422(6928), pp.198-207.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Banet, J.F., Billis, K., Girón, C.G., Hourlier, T. and Howe, K., 2016. The Ensembl gene annotation system. Database, 2016, p.baw093.

Anderson, N.G. and Anderson, N.L., 1996. Twenty years of two-dimensional electrophoresis: past, present and future. Electrophoresis, 17(3), pp.443-453.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. and Martin, M.J., 2004. UniProt: the universal protein knowledgebase. Nucleic acids research, 32(suppl 1), pp.D115-D119.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., 2000. Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), pp.25-29.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. and Martin, M.J., 2005. The universal protein resource (UniProt). Nucleic acids research, 33(suppl 1), pp.D154-D159.

Baldwin, M.A., 2004. Protein identification by mass spectrometry issues to be considered. Molecular & Cellular Proteomics, 3(1), pp.1-9.

Bantscheff, M., Lemeer, S., Savitski, M.M. and Kuster, B., 2012. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Analytical and bioanalytical chemistry, 404(4), pp.939-965.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. and Kuster, B., 2007. Quantitative mass spectrometry in proteomics: a critical review. Analytical and bioanalytical chemistry, 389(4), pp.1017-1031.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), pp.289-300.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. and Pilbout, S., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research, 31(1), pp.365-370.

Boeckmann, B., Blatter, M.C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B. and Bairoch, A., 2005. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. Comptes rendus biologies, 328(10), pp.882-899.

Bradshaw, R.A., Burlingame, A.L., Carr, S. and Aebersold, R., 2006. Reporting protein identification data the next generation of guidelines. Molecular & Cellular Proteomics, 5(5), pp.787-788.

Craig, R. and Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 20(9), pp.1466-1467.

Deutsch, E.W., Sun, Z., Campbell, D.S., Binz, P.A., Farrah, T., Shteynberg, D., Mendoza, L., Omenn, G.S. and Moritz, R.L., 2016. Tiered Human Integrated Sequence Search Databases for Shotgun Proteomics. Journal of Proteome Research.

Domon, B. and Aebersold, R., 2006. Challenges and opportunities in proteomics data analysis. Molecular & Cellular Proteomics, 5(10), pp.1921-1926.

Duncan, M.W., Aebersold, R. and Caprioli, R.M., 2010. The pros and cons of peptide-centric proteomics [AU: OK?]. Nature biotechnology, 28(7), p.1.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16), pp.3439-3440.

Durinck, S., Spellman, P.T., Birney, E. and Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols, 4(8), pp.1184-1191.

Elias, J.E. and Gygi, S.P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature methods, 4(3), pp.207-214.

Eng, J.K., Jahan, T.A. and Hoopmann, M.R., 2013. Comet: An open-source MS/MS sequence database search tool. Proteomics, 13(1), pp.22-24.

Eng, J.K., McCormack, A.L. and Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry, 5(11), pp.976-989.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. and Gil, L., 2011. Ensembl 2012. Nucleic acids research, p.gkr991.

Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl 1), pp.D258-D261.

Graves, P.R. and Haystead, T.A., 2002. Molecular biologist's guide to proteomics. Microbiology and Molecular Biology Reviews, 66(1), pp.39-63.

Griss, J., Martín, M., O'Donovan, C., Apweiler, R., Hermjakob, H. and Vizcaíno, J.A., 2011. Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB “complete proteome” sets. Proteomics, 11(22), pp.4434-4438.

Gupta, N. and Pevzner, P.A., 2009. False discovery rates of protein identifications: a strike against the two-peptide rule. Journal of proteome research, 8(9), pp.4173-4181.

Gygi, S.P. and Aebersold, R., 2000. Mass spectrometry and proteomics. Current opinion in chemical biology, 4(5), pp.489-494.

Gygi, S.P., Rist, B. and Aebersold, R., 2000. Measuring gene expression by quantitative proteome analysis. Current Opinion in Biotechnology, 11(4), pp.396-401.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S. and Coon, J.J., 2014. The one hour yeast proteome. Molecular & Cellular Proteomics, 13(1), pp.339-347.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S. and Spooner, W., 2016. Ensembl comparative genomics resources. Database, 2016, p.bav096.

Hsieh, E.J., Hoopmann, M.R., MacLean, B. and MacCoss, M.J., 2009. Comparison of database search strategies for high precursor mass accuracy MS/MS data. Journal of proteome research, 9(2), pp.1138-1143.

Huan, J., Hornick, N.I., Goloviznina, N.A., Kamimae-Lanning, A.N., David, L.L., Wilmarth, P.A., Mori, T., Chevillet, J.R., Narla, A., Roberts, C.T. and Loriaux, M.M., 2015. Coordinate regulation of residual bone marrow function by paracrine trafficking of AML exosomes. Leukemia.

Huang, T., Wang, J., Yu, W. and He, Z., 2012. Protein inference: a review. Briefings in bioinformatics, p.bbs004.

Kelleher, N.L., 2004. Peer reviewed: Top-down proteomics. Analytical chemistry, 76(11), pp.196-A.

Keller, A., Nesvizhskii, A.I., Kolker, E. and Aebersold, R., 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical chemistry, 74(20), pp.5383-5392.

Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. and Apweiler, R., 2004. The International Protein Index: an integrated database for proteomics experiments. Proteomics, 4(7), pp.1985-1988.

Keshishian, H., Burgess, M.W., Gillette, M.A., Mertins, P., Clauser, K.R., Mani, D.R., Kuhn, E.W., Farrell, L.A., Gerszten, R.E. and Carr, S.A., 2015. Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. Molecular & Cellular Proteomics, pp.mcp-M114.

Kessner, D., Chambers, M., Burke, R., Agus, D. and Mallick, P., 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics, 24(21), pp.2534-2536.

Koskinen, V.R., Emery, P.A., Creasy, D.M. and Cottrell, J.S., 2011. Hierarchical clustering of shotgun proteomics data. Molecular & Cellular Proteomics, 10(6), pp.M110-003822.

Liu, H., Sadygov, R.G. and Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Analytical chemistry, 76(14), pp.4193-4201.

Lundgren, D.H., Hwang, S.I., Wu, L. and Han, D.K., 2010. Role of spectral counting in quantitative proteomics. Expert review of proteomics, 7(1), pp.39-53.

Ma, K., Vitek, O. and Nesvizhskii, A.I., 2012. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. BMC bioinformatics, 13(Suppl 16), p.S1.

Ma, Z.Q., Dasari, S., Chambers, M.C., Litton, M.D., Sobecki, S.M., Zimmerman, L.J., Halvey, P.J., Schilling, B., Drake, P.M., Gibson, B.W. and Tabb, D.L., 2009. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. Journal of proteome research, 8(8), pp.3872-3881.

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K. and Horning, S., 2006. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Analytical chemistry, 78(7), pp.2113-2120.

Mann, M., Hendrickson, R.C. and Pandey, A., 2001. Analysis of proteins and proteomes by mass spectrometry. Annual review of biochemistry, 70(1), pp.437-473.

Matthiesen, R. ed., 2007. Mass spectrometry data analysis in proteomics (Vol. 1). Totowa, NJ: Humana Press.

McDonald, W.H. and Yates, J.R., 2002. Shotgun proteomics and biomarker discovery. Disease markers, 18(2), pp.99-105.

McDonald, W.H., Tabb, D.L., Sadygov, R.G., MacCoss, M.J., Venable, J., Graumann, J., Johnson, J.R., Cociorva, D. and Yates, J.R., 2004. MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Communications in Mass Spectrometry, 18(18), pp.2162-2168.

Nesvizhskii, A.I. and Aebersold, R., 2005. Interpretation of shotgun proteomic data the protein inference problem. Molecular & Cellular Proteomics, 4(10), pp.1419-1440.

Nesvizhskii, A.I., 2007. Protein identification by tandem mass spectrometry and sequence database searching. Mass Spectrometry Data Analysis in Proteomics, pp.87-119.

Nesvizhskii, A.I., 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of proteomics, 73(11), pp.2092-2123.

Nesvizhskii, A.I., Keller, A., Kolker, E. and Aebersold, R., 2003. A statistical model for identifying proteins by tandem mass spectrometry. Analytical chemistry, 75(17), pp.4646-4658.

Ong, S.E. and Mann, M., 2005. Mass spectrometry–based proteomics turns quantitative. Nature chemical biology, 1(5), pp.252-262.

Pages, H., Gentleman, R., Aboyoun, P. and DebRoy, S., Biostrings: String objects representing biological sequences, and matching algorithms, 2008. R package version, 2(0), p.160.

Pandey, A. and Mann, M., 2000. Proteomics to study genes and genomes. Nature, 405(6788), pp.837-846.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18), pp.3551-3567.

Pruitt, K.D., Tatusova, T. and Maglott, D.R., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research, 33(suppl 1), pp.D501-D504.

Rappsilber, J. and Mann, M., 2002. What does it mean to identify a protein in proteomics? Trends in biochemical sciences, 27(2), pp.74-78.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S. and Purkayastha, S., 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics*, *3*(12), pp.1154-1169.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.

Searle, B.C., 2010. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics, 10(6), pp.1265-1269.

Senko, M.W., Remes, P.M., Canterbury, J.D., Mathur, R., Song, Q., Eliuk, S.M., Mullen, C., Earley, L., Hardman, M., Blethrow, J.D. and Bui, H., 2013. Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. Analytical chemistry, 85(24), pp.11710-11714.

Serang, O., MacCoss, M.J. and Noble, W.S., 2010. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. Journal of proteome research, 9(10), pp.5346-5357.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C., 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry*, *75*(8), pp.1895-1904.

Wilmarth, P.A., Riviere, M.A. and David, L.L., 2009. Techniques for accurate protein identification in shotgun proteomic studies of human, mouse, bovine, and chicken lenses. Journal of ocular biology, diseases, and informatics, 2(4), pp.223-234.

Yates III, J.R., Park, S.K.R., Delahunty, C.M., Xu, T., Cociorva, D. and Carvalho, P.C., 2012. Toward objective evaluation of proteomic algorithms. Nature methods, 9(5), p.455.

Yates, J.R. and Kelleher, N.L., 2013. Top down proteomics. Anal Chem, 85(13), p.6151.

Yates, J.R., 2000. Mass spectrometry: from genomics to proteomics. Trends in Genetics, 16(1), pp.5-8.

Yates, J.R., Ruse, C.I. and Nakorchevsky, A., 2009. Proteomics by mass spectrometry: approaches, advances, and applications. Annual review of biomedical engineering, 11, pp.49-79.

Zhang, B., Chambers, M.C. and Tabb, D.L., 2007. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. Journal of proteome research, 6(9), pp.3549-3557.

Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. and Yates III, J.R., 2013. Protein analysis by shotgun/bottom-up proteomics. Chemical reviews, 113(4), pp.2343-2394.

Figures

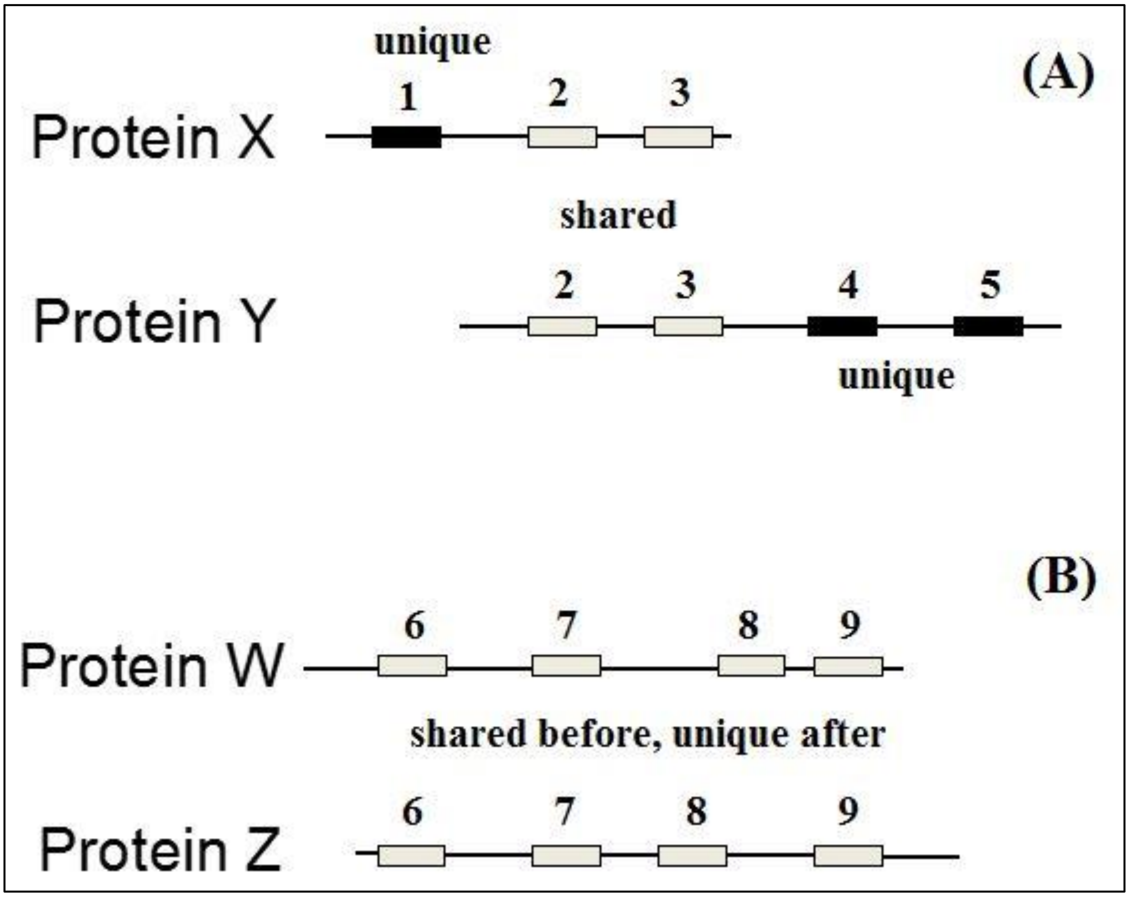

Figure 1.1. Illustration of distinguishable/indistinguishable proteins. Definitions of shared and unique peptides for distinguishable (A) and indistinguishable proteins (B). Before protein inference, shared peptides are those that map to multiple proteins in the protein database, and unique peptides map to only single protein database entries. Both proteins X and Y in (A) are inferred to be present in the sample due to the unique peptide evidence. In (B) there is no evidence to distinguish protein W from protein Z, so a redundant protein group containing both proteins is inferred to be present in the sample. There is ambiguity associated with proteins W and Z. After protein inference, definitions of shared and unique are redefined with respect to the list of inferred proteins instead of the original protein database. The shared peptides associated with proteins W and Z are redefined as unique to the protein group (if they are not associated with any other protein/protein groups).

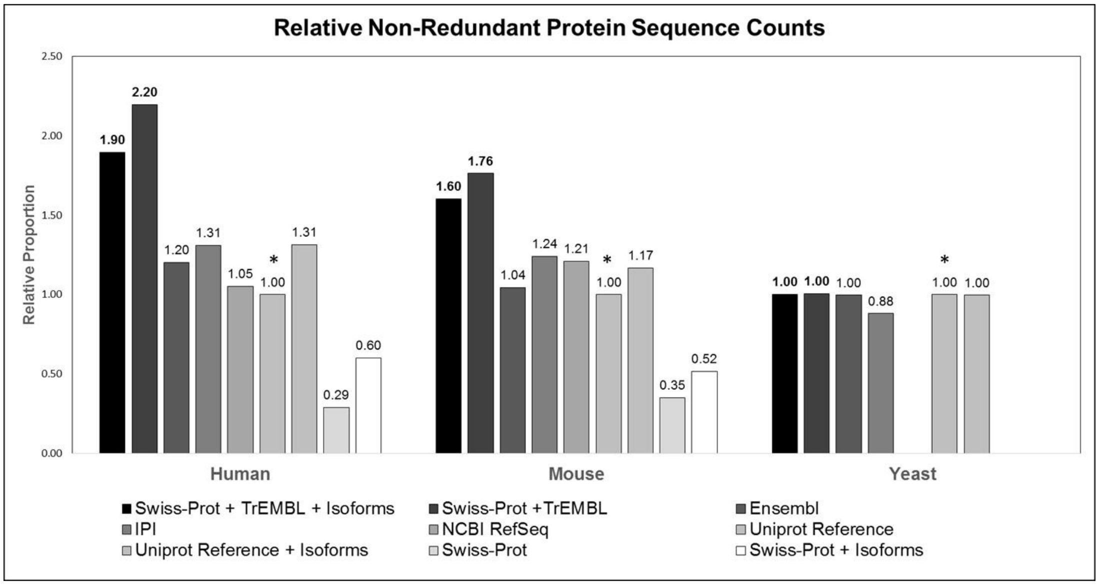

Figure 3.1. **Relative non-redundant protein sequence counts.** Relative non-redundant protein counts for human, mouse, and yeast as a function of protein database. Protein database sizes are relative to the UniProt Reference proteome for each organism. The UniProt reference proteomes are indicated with asterisks (*).

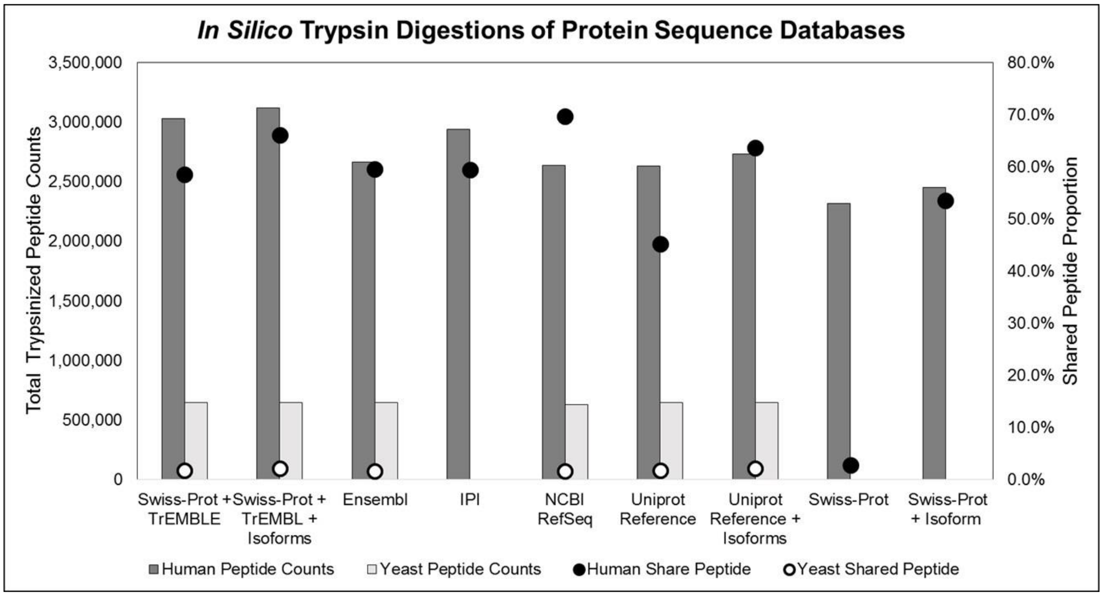

Figure 3.2. *In silico* trypsin digestions of sequence databases. Comparison of *in silico* digestion sizes between protein databases. Dark grey columns indicate total peptide counts generated from *in silico* tryptic digestions of non-redundant proteins from human sequence databases. Black circles represent the percentage of shared peptides in human protein sequence databases. Light grey columns indicate total peptide counts generated from *in silico* tryptic digestions of non-redundant proteins from yeast sequence databases. White circles represent the percentage of shared peptides in yeast protein sequence databases.

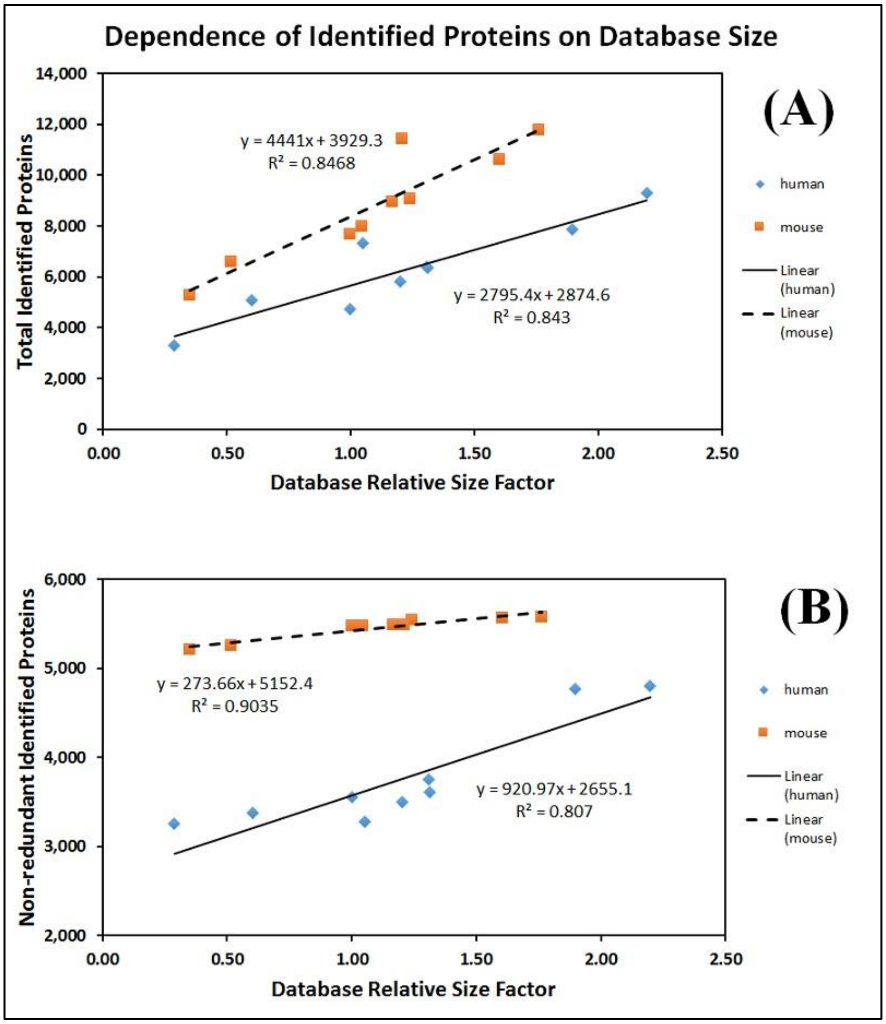

Figure 3.3. Dependence of identified proteins on database size. Dependence of the number of identified proteins for human and mouse on the size of the protein database. The trend lines for total protein identifications (redundant counts) in (A) have much larger slopes do the trend lines for non-redundant protein identifications (B) indicating a greater dependence on database size.

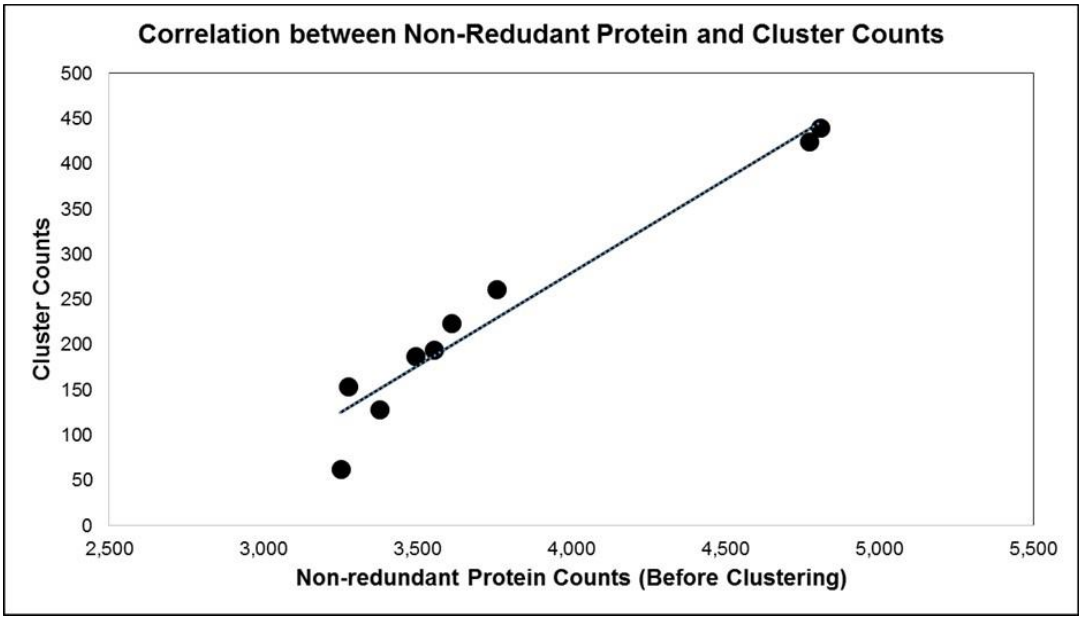

Figure 3.4. Correlation between non-redundant proteins and cluster counts. Correlation between non-redundant protein counts (prior to clustering) and the cluster sizes after processing the human plasma sample results with the PAW clustering algorithm.

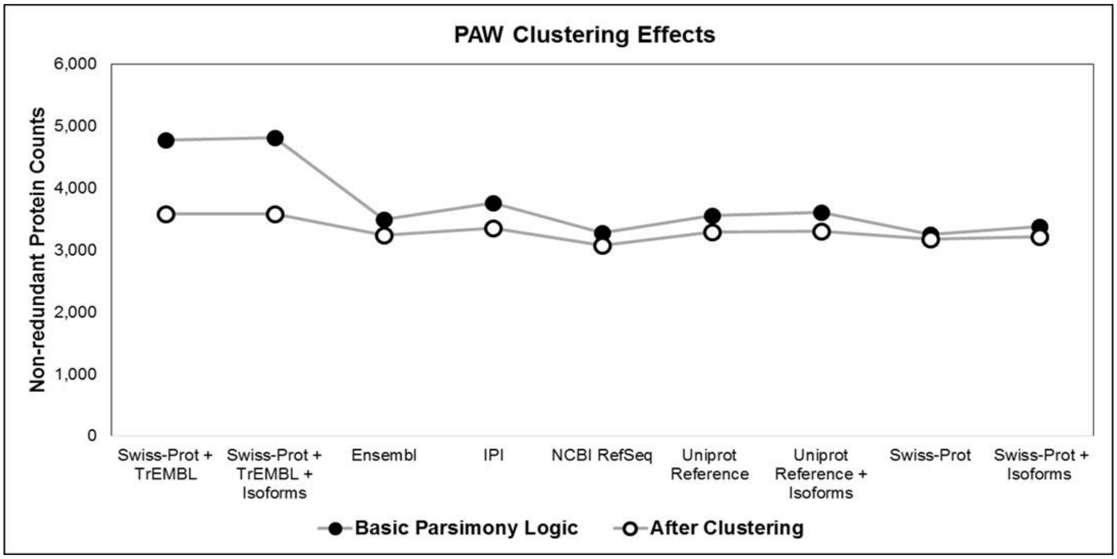

Figure 3.5. PAW clustering effects. The effects of PAW clustering on non-redundant protein counts in human plasma samples. Dark circles represent non-redundant protein counts before and white circles represent non-redundant protein counts after implementing PAW clustering.

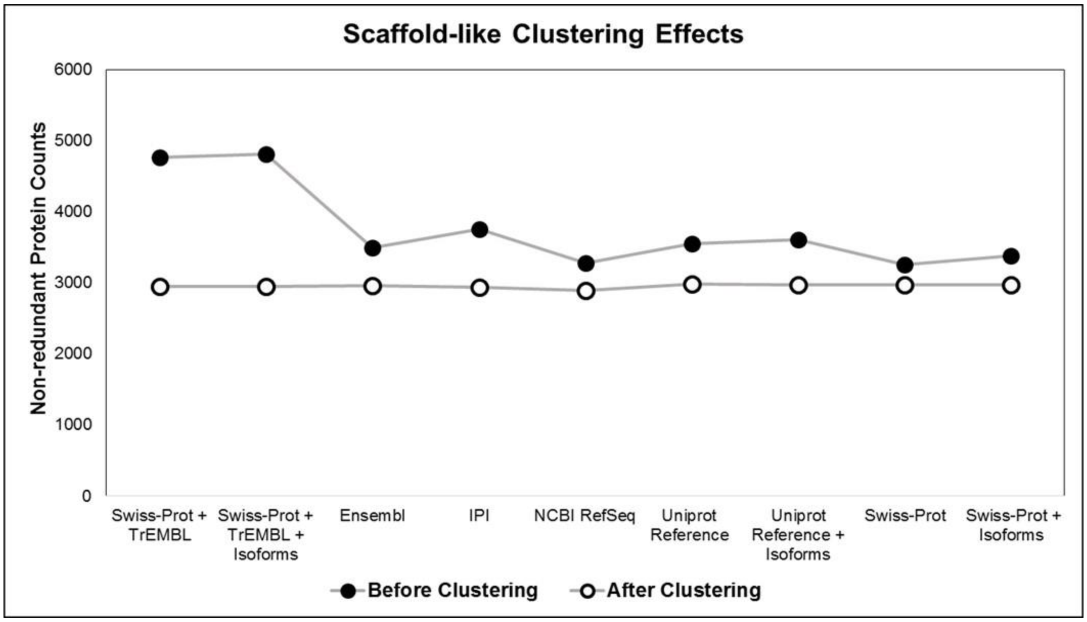

Figure 3.6. Scaffold-like clustering effects. The results of Scaffold-like clustering on non-redundant protein counts in human plasma samples. Dark circles represent non-redundant protein counts before and white circles represent non-redundant protein counts after implementing Scaffold-like clustering.

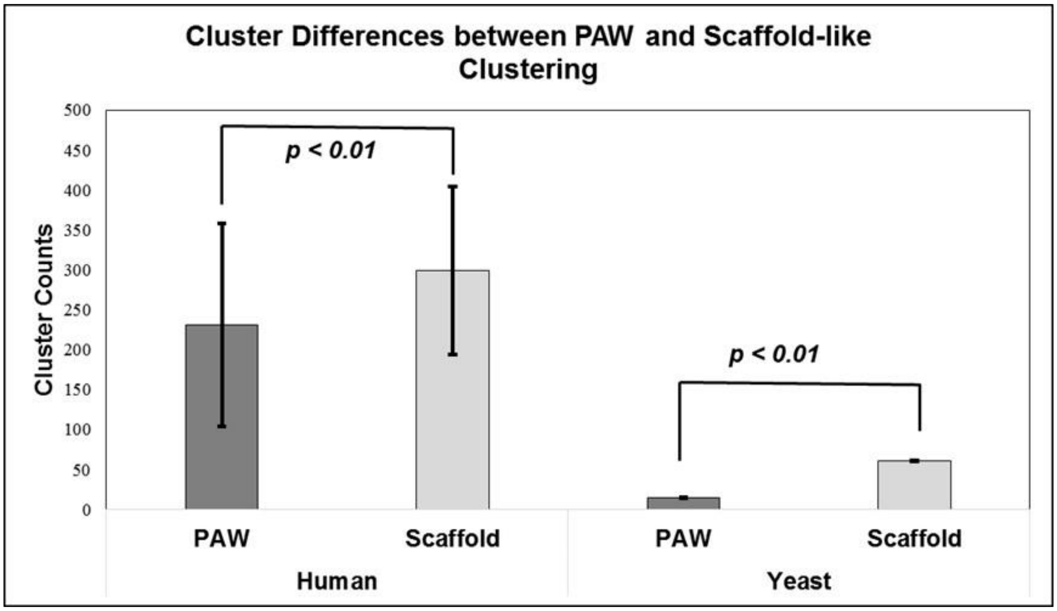

Figure 3.7. Cluster differences between PAW and Scaffold-like algorithms. The differences in numbers of clusters between PAW and Scaffold-like clustering algorithms for human and yeast biological samples. Dark grey columns represent average cluster counts from all databases in human and yeast biological samples using the PAW clustering algorithm. Light grey columns represent average cluster counts from all protein databases in human and yeast biological samples using the Scaffold-like clustering algorithm. Scaffold-like clustering had a significantly greater number of clusters than PAW clustering for both human and yeast. Two-sided paired t-test was performed to determine significance.

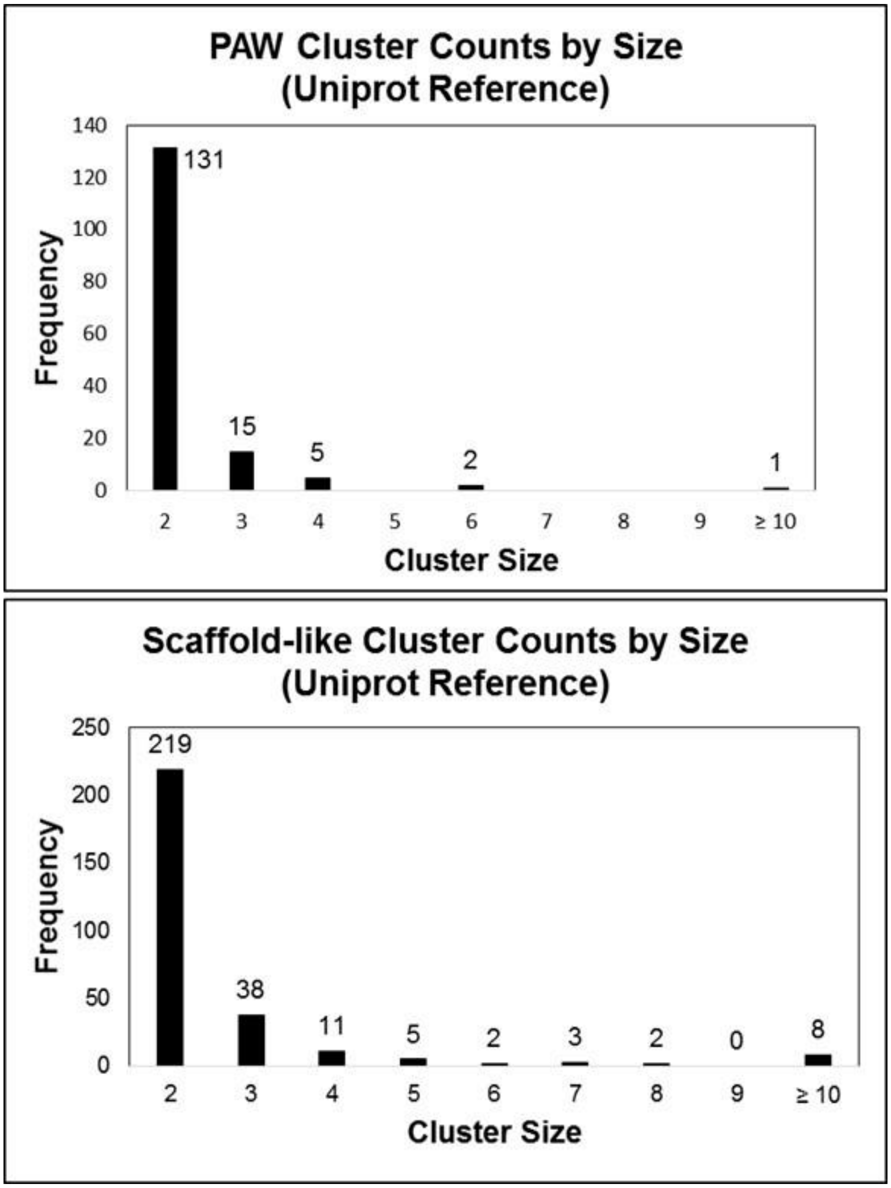

Figure 3.8. Frequency of cluster sizes. Histograms of cluster sizes identified in human plasma samples searched against the UniProt Reference canonical proteome for PAW (top) and Scaffold-like (bottom) clustering algorithms. Scaffold-like clustering resulted in more clusters and a greater variety of cluster sizes than did the PAW clustering.

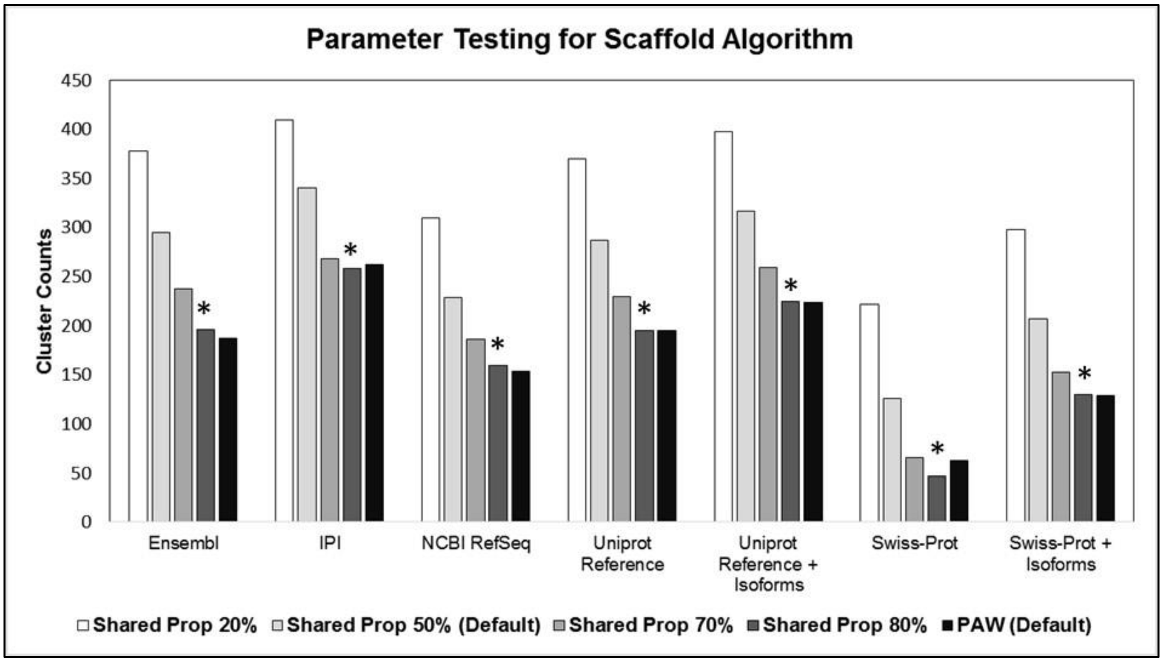

Figure 3.9. Testing of parameters for Scaffold-like algorithm. Changes in the number of clusters generated with different shared peptide proportion thresholds (20%, 50%, 70%, and 80%) in Scaffold-like clustering. Black columns represent the clusters generated by PAW clustering algorithm at its default setting (Pseudo spectral count of 2.0, Shared spectral threshold of 20.0, Fraction of shared to unique spectral count threshold of 40.0). An 80% shared peptide threshold for Scaffold-like clustering (indicated by '*') produced similar numbers of clusters as the PAW clustering algorithm.

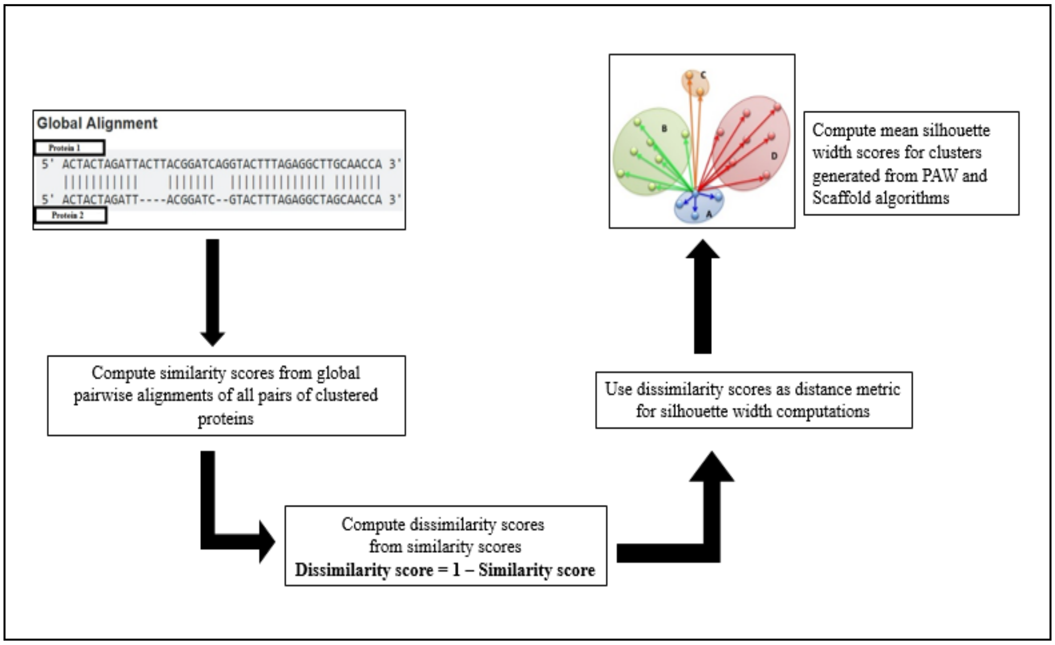
Global Alignment
Protein 1
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Protein 2
Compute similarity scores from global pairwise alignments of all pairs of clustered proteins
Compute dissimilarity scores from similarity scores
Dissimilarity score = 1 – Similarity score
Use dissimilarity scores as distance metric for silhouette width computations
C
B
D
A
Compute mean silhouette width scores for clusters generated from PAW and Scaffold algorithms

Figure 3.10. Illustration of silhouette score computation. Internal cluster evaluation procedure for computing the silhouette scores for clusters generated from human and yeast samples using either PAW or Scaffold-like algorithms.

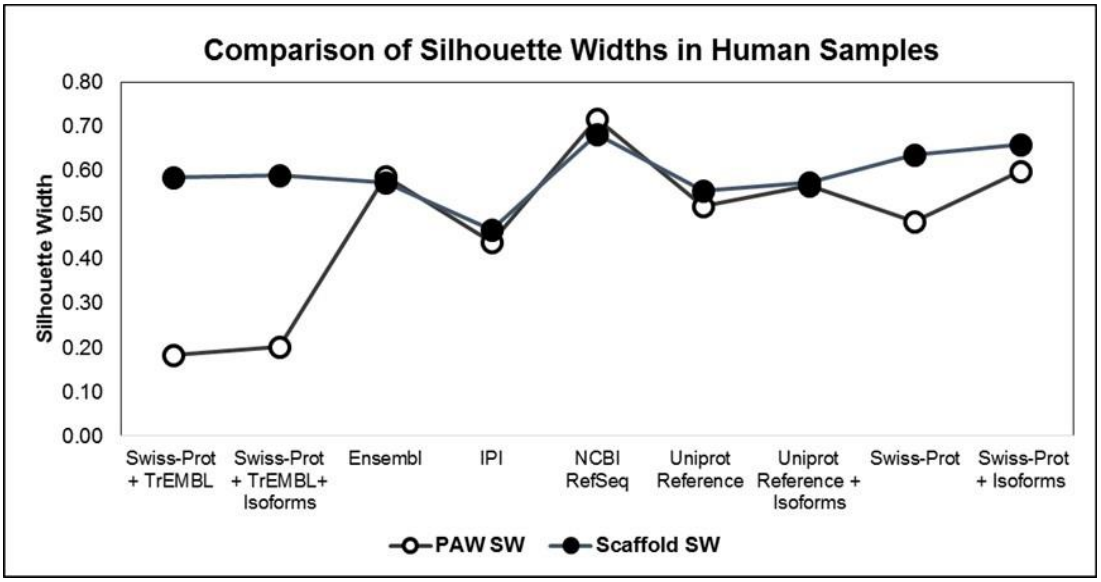

Figure 3.11. Comparison of silhouette scores between PAW and Scaffold-like algorithms. Mean silhouette scores using PAW and Scaffold-like algorithms on human samples across all sequence databases. Black circles indicate the mean silhouette scores from Scaffold-like clustering. White circles indicate the mean silhouette scores from PAW clustering.

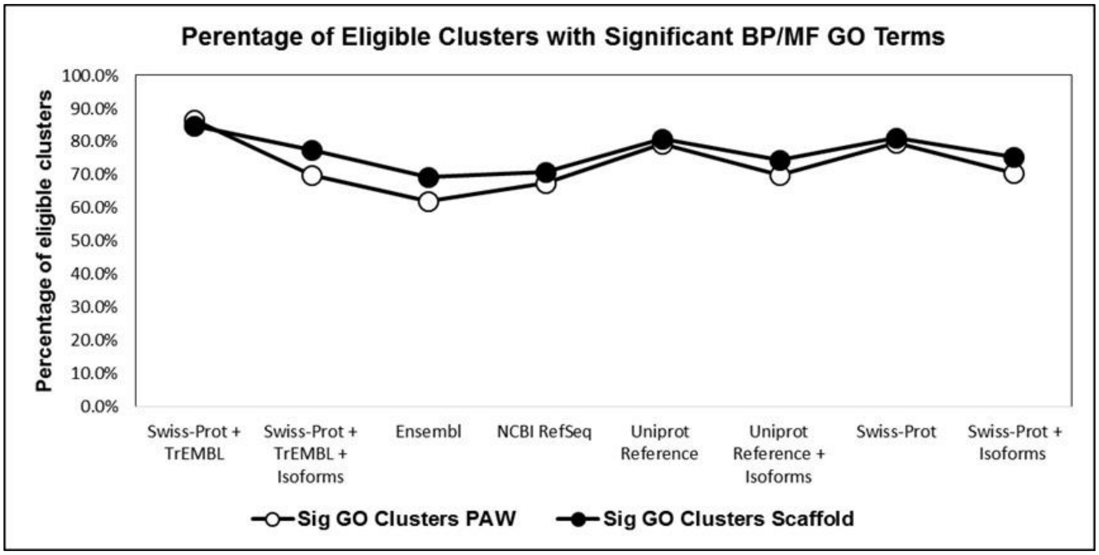

Figure 3.12. Cluster validation using gene enrichment analyses. Percentages of eligible clusters generated from PAW and Scaffold-like algorithms in human plasma samples with significant BP or MF GO terms. Black circles indicate the proportion of PAW eligible clusters that have either significant BP or MF GO terms. White circles indicate the proportion of eligible Scaffold-like clusters that have either significant BP or MF GO terms.

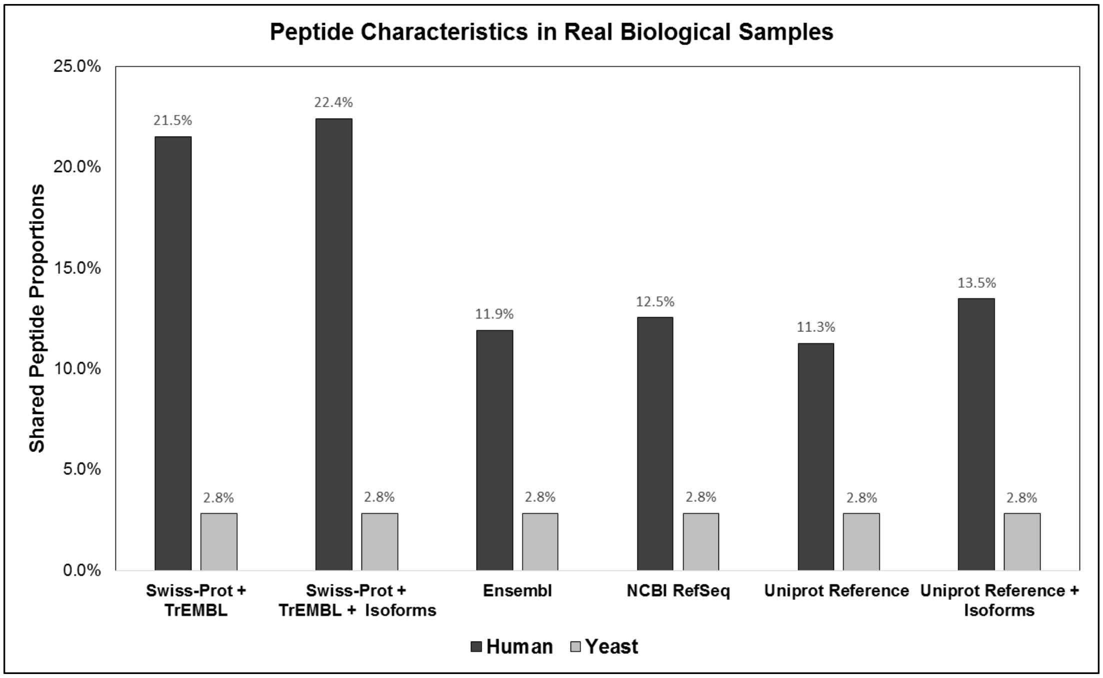

Figure 3.13. Peptide characteristics in real biological samples. The proportion of shared peptides in human plasma samples and yeast whole cell lysates are shown. Shared and unique peptides were defined in the context of the basic parsimonious protein list.

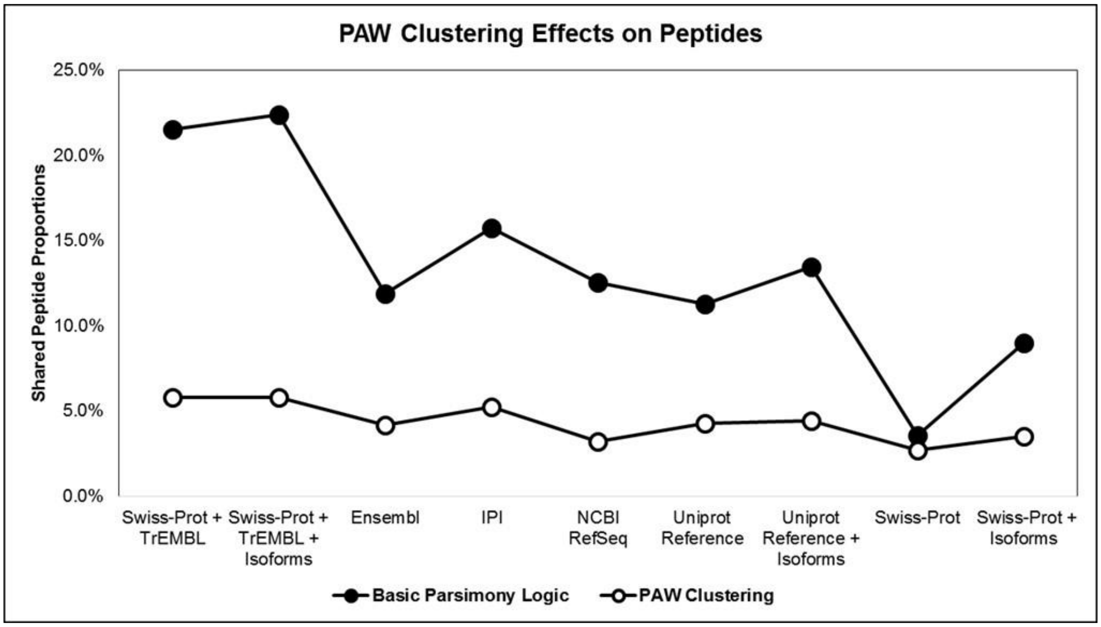

Figure 3.14. PAW clustering effects on peptide characteristics. The results of PAW clustering on peptide counts in human plasma samples. Dark circles represent shared peptide proportions before and white circles represent shared peptide proportions after implementing PAW clustering.

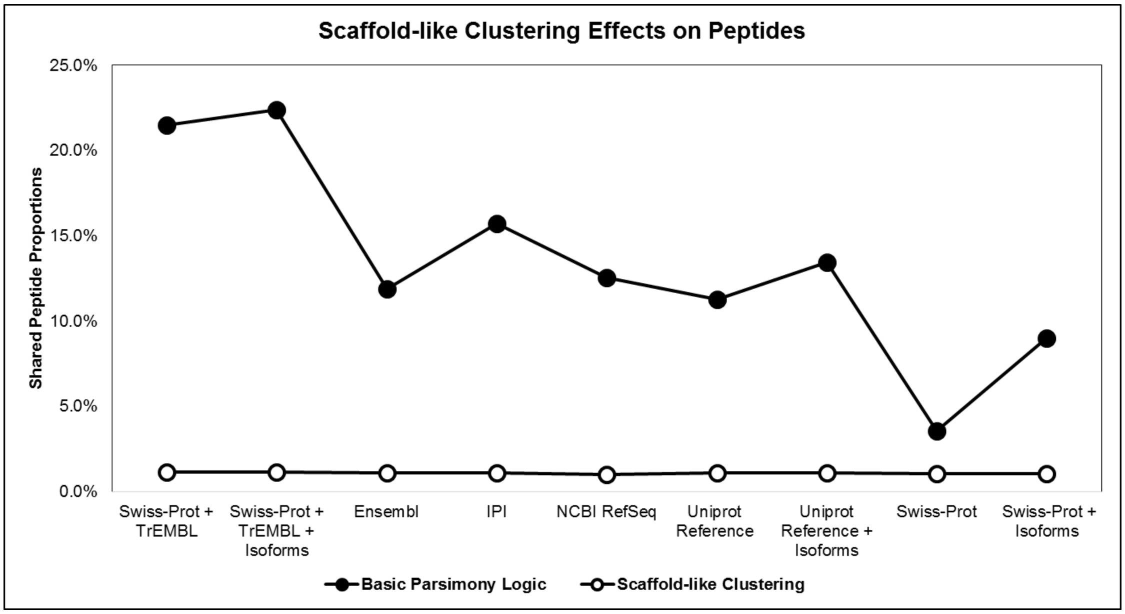

Figure 3.15. Scaffold-like clustering effects on peptide characteristics. The results of Scaffold-like clustering on peptide counts in human plasma samples. Dark circles represent shared peptide proportions before and white circles represent shared peptide proportions after implementing Scaffold-like clustering.

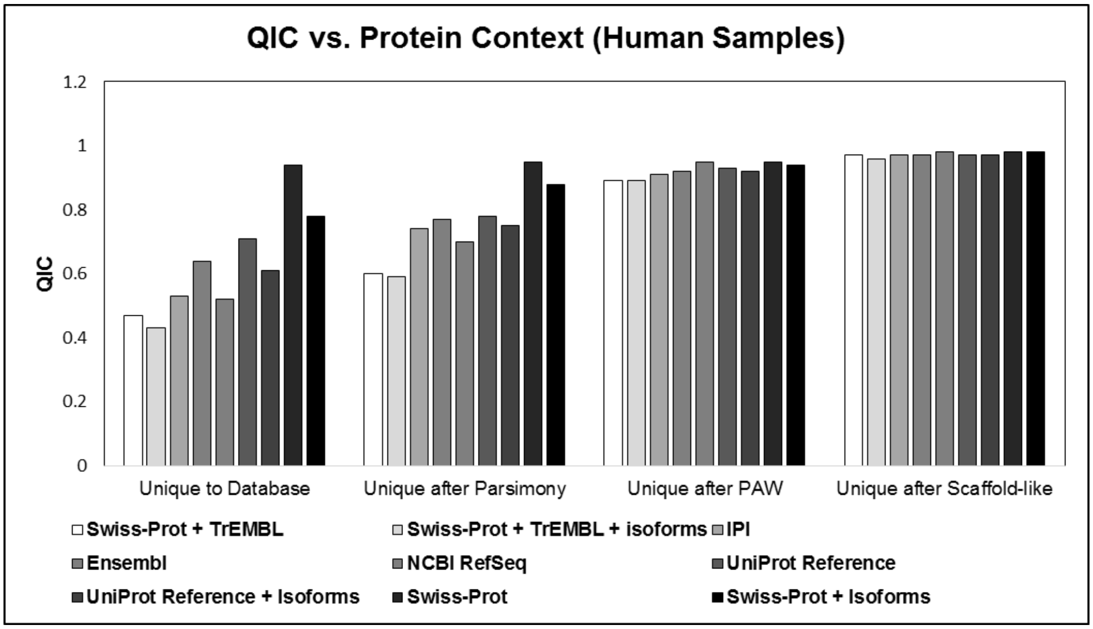

Figure 3.16. Change in QIC with changing protein context. Data searched against the nine protein sequence database are shown. QIC was calculated in three different protein contexts (protein database, basic parsimony, and extended parsimony). Both PAW clustering and Scaffold-like clustering results are shown.

Tables

Table 2.1. Sources of protein sequence databases.

Databases denoted by Swiss-Prot indicate canonical sequences only unless “isoforms” is explicitly stated. All databases were downloaded from the Internet on February 2016. The download procedure varied depending on the database source and there can be more than one way to retrieve protein databases. Ensembl and IPI maintain FTP sites for FASTA file. NCBI and Uniprot have both FTP mechanisms and direct download options; direct download were used when possible. Uniprot uses URL syntax to specify download options, NCBI does not.

Database	Version	Download Link[1]
Swiss-Prot + TrEMBL	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=taxonomy:9606&format=fasta&include=no
Swiss-Prot + TrEMBL + Isoforms	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=taxonomy:9606&format=fasta&include=yes
IPI[2]	3.87	ftp://ftp.ebi.ac.uk/pub/databases/IPI/last_release/current/ipi.HUMAN.fasta.gz
Ensembl	83	ftp:://ftp.ensembl.org/pub/release-83/pep/Homo_sapiens.GRCh38.pep.all.fa.gz
NCBI RefSeq	75	http://www.ncbi.nlm.nih.gov/protein {txid9606[Organism:noexp] AND refseq[filter] & Send to File in FASTA format}[3]
UniProt Reference	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=proteome:up000005640&format=fasta&include=no
UniProt Reference + Isoforms	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=proteome:up000005640&format=fasta&include=yes
Swiss-Prot[2]	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=taxonomy:9606&fil=reviewed:yes&format=fasta&include=no
Swiss-Prot + Isoforms[2]	2016.01	http://www.uniprot.org/uniprot/?sort=&desc=&compress=yes&query=taxonomy:9606&fil=reviewed:yes&format=fasta&include=yes

[1] Link examples are for human (taxon=9606); mouse and yeast would be similar except for taxon number (10090 and 559292).

[2] These databases were not used for yeast. There is no IPI database for yeast and there were so few TrEMBL sequences for yeast that Swiss-Prot + TrEMBL is essentially redundant with Swiss-Prot.

[3] The NCBI database selection and download process is interactive so a download link is not possible.

Table 3.1. Total human, mouse, and yeast protein sequence counts from sequence databases.

Multiple copies of 100% identical sequences for a protein in a species are considered duplicate sequences. Non-redundant protein databases generated after removal of duplicate sequences were used for theoretical digests and in the database searches.

	Human		Mouse		Yeast	
Database	**Total Sequence Count**	**Duplicate Sequence Count**	**Total Sequence Count**	**Duplicate Sequence Count**	**Total Sequence Count**	**Duplicate Sequence Count**
Swiss-Prot + TrEMBL	150,227	17,829	79,950	3,111	6,729	84
Swiss-Prot + TrEMBL + Isoforms	172,164	18,699	88,002	3,519	6,751	86
IPI[1]	91,464	0	59,534	0		
Ensembl	102,450	18,456	56,999	6,963	6,692	82
NCBI RefSeq	100,408	26,955	78,310	20,371	5,907	63
UniProt Reference	69,986	124	50,189	2,260	6,721	84
UniProt Reference + Isoforms	91,923	185	58,239	2,289	6,743	86
Swiss-Prot[1]	20,187	41	16,761	5		
Swiss-Prot + Isoforms[1]	42,124	43	24,813	5		

[1] These databases were not used for yeast.

Table 3.2. Tryptic peptide counts in human, mouse, and yeast databases.

In silico tryptic digestions of the protein databases were performed and the total number of distinct peptide sequences (digestion size) are shown in the first columns for each species. Peptide sequences liberated from a single protein database entry were classified as unique, and peptides originating from multiple proteins were classified as shared (degenerate).

	Human		Mouse		Yeast	
Database	Total Distinct Peptide Count	Total Shared Peptide Count	Total Distinct Peptide Count	Total Shared Peptide Count	Total Distinct Peptide Count	Total Shared Peptide Count
Swiss-Prot + TrEMBL	3,029,415	1,771,254	2,742,886	1,583,750	647,816	11,293
Swiss-Prot + TrEMBL + Isoforms	3,116,511	2,057,760	2,774,249	1,777,097	647,852	13,769
IPI[1]	2,940,847	1,746,490	2,694,266	1,381,986		
Ensembl	2,665,253	1,585,695	2,492,844	1,067,312	646,533	11,048
NCBI RefSeq	2,636,739	1,836,420	2,586,741	1,607,850	630,314	10,835
UniProt Reference	2,629,090	1,188,164	2,498,296	1,042,427	647,667	11,288
UniProt Reference + Isoforms	2,730,931	1,738,326	2,533,469	1,327,175	647,703	13,764
Swiss-Prot[1]	2,315,554	63,673	1,965,275	35,427		
Swiss-Prot + Isoforms[1]	2,452,596	1,312,783	2,012,492	647,201		

[1] These databases were not used for yeast.

Table 3.3. Protein characteristics of human, mouse, and yeast samples after basic parsimony logic.

The PAW pipeline, the two peptide rule, and basic parsimony principles were used for confident protein identifications.

	Human			Mouse			Yeast		
Database	**Total Protein Count**	**NR[2] Protein Count**	**Number of Groups[3]**	**Total Protein Count**	**NR[2] Protein Count**	**Number of Groups[3]**	**Total Protein Count**	**NR[2] Protein Count**	**Number of Groups[3]**
Swiss-Prot + TrEMBL	7,858	4,771	1,541	10,622	5,571	2,609	3,811	3,782	23
Swiss-Prot + TrEMBL + Isoforms	9,296	4,806	1,878	11,760	5,574	2,995	3,824	3,783	34
IPI[1]	6,404	3,757	1,295	9,053	5,544	2,023			
Ensembl	5,811	3,494	1,181	7,966	5,472	1,509	3,807	3,779	22
NCBI RefSeq	7,341	3,277	1,405	11,413	5,488	2,283	3,811	3,783	22
UniProt Reference	4,749	3,553	732	7,679	5,473	1,436	3,811	3,782	23
UniProt Reference + Isoforms	6,363	3,611	1,322	8,952	5,489	1,946	3,811	3,782	34
Swiss-Prot[1]	3,314	3,251	40	5,261	5,216	30			
Swiss-Prot + Isoforms[1]	5,089	3,376	978	6,604	5,253	923			

[1] These databases were not used for yeast.

[2] Non-Redundant Protein Count.

[3] Number of Groups = Number of Clusters + Number of Identified Redundant Groups.

Table 3.4 Step-by-step trace of the protein inference process in the PAW pipeline for human samples.

The complete mapping of all identified peptides to respective proteins is followed by removing any peptide sets that do not have at least two members. Peptide sets are compared and combined if they are identical, or removed if entirely contained in another peptide set. The final culling is to remove proteins with insufficient evidence per experimental sample.

Database	All Mapped Proteins	After Two Peptides per Protein	After Identical Peptide Set Grouping	After Peptide Subset Removal	Non-Redundant Proteins per Sample[1]
Swiss-Prot + TrEMBL	55,282	31,975	15,816	5,839	5,087
Swiss-Prot + TrEMBL + Isoforms	65,593	36,175	16,982	5,894	5,135
IPI	35,881	15,947	10,175	4,792	4,087
Ensembl	33,871	15,032	9,721	4,618	3,856
NCBI RefSeq	40,568	16,686	7,276	4,380	3,625
UniProt Reference	26,514	11,891	8,889	4,666	3,903
UniProt Reference + Isoforms	37,862	16,649	10,254	4,766	3,982
Swiss-Prot	11,238	4,838	4,691	4,416	3,632
Swiss-Prot + Isoforms	22,972	9,656	6,324	4,544	3,763

[1] Includes contaminants and decoys.

Table 3.5. Effects of extended parsimony clustering on protein counts for human, mouse, and yeast samples.

The basic parsimony, PAW clustering and Scaffold-like clustering non-redundant protein counts are shown for human, mouse, and yeast samples from the different species-specific databases.

	Human			Mouse			Yeast		
Database	**Basic Parsimony**	**PAW Clustering**	**Scaffold-like Clustering**	**Basic Parsimony**	**PAW Clustering**	**Scaffold-like Clustering**	**Basic Parsimony**	**PAW Clustering**	**Scaffold-like Clustering**
Swiss-Prot + TrEMBL	4,771	3,580	2,943	5,571	5,457	5,173	3,782	3,744	3,681
Swiss-Prot + TrEMBL + Isoforms	4,806	3,583	2,950	5,574	5,451	5,175	3,783	3,744	3,681
IPI[1]	3,757	3,355	2,942	5,544	5,443	5,181			
Ensembl	3,494	3,238	2,964	5,472	5,400	5,176	3,779	3,742	3,678
NCBI RefSeq	3,277	3,074	2,892	5,488	5,403	5,180	3,783	3,744	3,681
UniProt Reference	3,553	3,293	2,981	5,473	5,409	5,185	3,782	3,744	3,681
UniProt Reference + Isoforms	3,611	3,296	2,976	5,489	5,411	5,183	3,782	3,744	3,681
Swiss-Prot[1]	3,251	3,170	2,970	5,216	5,193	5,020			
Swiss-Prot + Isoforms[1]	3,376	3,209	2,976	5,253	5,206	5,020			

[1] These databases were not used for yeast.

Table 3.6. Summary statistics of clusters generated after extended parsimony clustering.

Cluster characteristics for human plasma and yeast whole cell lysate samples are computed after implementing PAW clustering and Scaffold-like clustering algorithms. Singleton clusters were not included when calculating cluster mean sizes. Minimum size of clusters is always 2.

Database	PAW Clustering						Scaffold-like Clustering					
	Human			Yeast			Human			Yeast		
	Cluster Count	Mean Size	Max Size	Cluster Count	Mean Size	Max Size	Cluster Count	Mean Size	Max Size	Cluster Count	Mean Size	Max Size
Swiss-Prot + TrEMBL	425	3.8	91	14	3.7	20	442	5	785	60	2.7	30
Swiss-Prot + TrEMBL + Isoforms	440	3.7	79	15	3.6	20	451	5	783	61	2.7	30
IPI[1]	262	2.5	10				340	3.3	126			
Ensembl	187	2.3	6	15	3.5	20	295	2.7	31	61	2.7	29
NCBI RefSeq	154	2.3	9	15	3.6	20	229	2.6	28	62	2.6	30
UniProt Reference	195	2.3	6	14	3.7	20	287	2.9	58	60	2.7	30
UniProt Reference + Isoforms	224	2.3	8	14	3.7	20	317	2.9	58	60	2.7	30
Swiss-Prot[1]	63	2.2	5				126	3	54			
Swiss-Prot + Isoforms[1]	129	2.2	7				207	2.8	54			

[1] These databases were not used for yeast.

Table 3.7. Summary statistics of clusters used in the GO enrichment external cluster validation.

Large fractions of proteins within clusters generated by PAW and Scaffold-like clustering algorithms did not have unique Entrez gene IDs and could not be analyzed. Eligible clusters contain at least two unique Entrez Gene IDs. Uncorrected p-values were computed for each GO term present within a cluster using Fisher Exact Test and further corrected for multiple testing using Benjamini/Hochberg FDR method with a restriction at 0.05. Significantly enriched clusters are those clusters that have at least one GO term with an adjusted p-value of less than 0.05 that is associated with all the members of the corresponding protein cluster.

	PAW Clustering						**Scaffold-like Clustering**					
Database	**Cluster Count**	**Proteins in Clusters**	**Proteins with Entrez Gene ID**	**Proteins with Unique Entrez Gene ID**	**Eligible Cluster Count**	**Significantly Enriched Clusters[1]**	**Cluster Count**	**Proteins in Clusters**	**Proteins with Entrez Gene ID**	**Proteins with Unique Entrez Gene ID**	**Eligible Cluster Count**	**Significantly Enriched Clusters[1]**
Swiss-Prot + TrEMBL	425	1601	234	234	15	13/13	442	2229	377	376	43	36/37
Swiss-Prot + TrEMBL + Isoforms	440	1646	321	270	20	14/14	451	2264	492	418	47	36/37
Ensembl	187	430	387	223	46	28/29	295	792	702	448	111	74/80
NCBI	154	348	336	202	46	31/31	229	586	567	389	108	74/79
Uniprot Reference	195	443	192	179	29	23/23	287	827	385	339	65	51/54
Uniprot Reference + Isoforms	224	526	296	217	30	21/21	317	919	521	383	71	51/55
Swiss-Prot	63	136	117	101	42	33/34	126	380	331	276	109	85/92
Swiss-Prot + Isoforms	129	287	267	169	44	31/31	207	579	531	361	112	82/87

[1] Biological Process/Molecular Function.

Table 3.8. Changes in QIC with respect to different protein contexts.

The different contexts are: PSMs unique with respect to the protein sequence database, PSMs unique with respect to the parsimonious list of identified proteins, and PSMs unique with respect to the refined list of identified proteins after extended parsimony clustering (both PAW and Scaffold-like clustering are shown).

Database	Biological Sample	Total PSMs	Unique to Protein Database	Unique After Basic Parsimony	Unique After PAW Clustering	Unique After Scaffold-like Clustering
Swiss-Prot + TrEMBL + Isoforms	Human	826,483	351,380	489,520	732,145	797,403
Swiss-Prot + TrEMBL	Human	828,870	388,744	497,427	735,501	799,968
IPI	Human	816,653	434,003	603,387	744,720	789,577
Ensembl	Human	815,889	522,155	625,506	753,505	789,926
NCBI RefSeq	Human	795,896	416,099	553,354	754,683	778,682
UniProt Reference + Isoforms	Human	817,730	500,456	612,228	753,846	791,216
UniProt Reference	Human	817,222	576,659	636,288	756,462	791,039
Swiss-Prot + Isoforms	Human	807,899	629,048	708,937	763,311	791,066
Swiss-Prot	Human	806,264	761,626	762,634	768,142	789,719
Swiss-Prot + TrEMBL + Isoforms	Mouse	85,709	38,733	74,749	78,448	82,589
Swiss-Prot + TrEMBL	Mouse	85,662	42,424	75,345	78,380	82,565
IPI	Mouse	85,454	52,747	75,565	78,888	82,442
Ensembl	Mouse	85,200	60,868	76,667	79,079	82,203
NCBI RefSeq	Mouse	85,220	51,385	76,704	79,419	82,281
UniProt Reference + Isoforms	Mouse	85,530	56,127	76,824	79,335	82,487
UniProt Reference	Mouse	85,551	62,123	77,603	79,404	82,511
Swiss-Prot + Isoforms	Mouse	82,872	66,223	75,909	77,518	80,095
Swiss-Prot	Mouse	82,838	76,458	77,094	77,579	80,068
Swiss-Prot + TrEMBL + Isoforms	Yeast	184,050	172,448	174,291	175,202	180,505
Swiss-Prot + TrEMBL	Yeast	184,006	173,644	174,483	175,160	180,461
Ensembl	Yeast	183,923	173,512	174,351	175,084	180,391
NCBI RefSeq	Yeast	184,229	173,573	174,416	175,370	180,684
UniProt Reference + Isoforms	Yeast	184,025	172,660	174,502	175,179	180,482
UniProt Reference	Yeast	184,006	173,644	174,483	175,160	180,461