INTEGRATED DATA MODELING IN

HIGH-THROUGHPUT PROTEOMICES

By

SHUANGSHUANG JIN

A dissertation submitted in partial fulfillment of
The requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

DECEMBER 2007

To the Faculty of Washington State University:

The members of the Committee appointed to examine the Dissertation of SHUANGSHUANG JIN find it satisfactory and recommend that it be accepted.

_____

Chair

_____

_____

ACKNOWLEDGMENT

The work of this dissertation has been an inspiring, exciting, interesting, and challenging experience. It cannot be accomplished without the accompanying and support by many people.

The first person I would like to express my deep and sincere gratitude to is my direct supervisor, Dr. John H. Miller, who has given me countless suggestions and insightful guidance during the course of this work. His wide knowledge and fruitful discussions have been great value to me. His understanding, encouragement and instruction have provided a good basis for the present dissertation. He's definitely one of my great supervisors in my life.

I am deeply grateful to my committee members, Dr. Donald J. Lynch, and Dr. Robert R. Lewis. They accompanied me through each challenging examination: qualification exam, preliminary exam and the final oral defense of this dissertation. Without their understanding and support, I may not able to get the chance to fulfill my academic goal step by step. Their insightful comments and suggestions on this thesis work have also been extremely helpful. The experience of working with Dr. Lewis on my Master's thesis was also an unforgettable exciting thing in my life.

I wish to express my warm thanks to the colleagues in Pacific Northwest National Laboratory: Dr. David Springer, Mrs. Renee Johnson, and Dr. Don Dally, etc. It was my great pleasure to work with them. We also gratefully acknowledge the grant support by the Office of Science (BER), U. S. Department of Energy.

Many thanks to WSU Graduate School and EECS for providing me such a good environment to accomplish my graduate program. The support from EECS administration: Dr. Ali Saberi and Dr. Zhe Dang (Pullman), Dr. Donald Lynch and Dr. Robert Lewis (Tri-Cities); the help from EECS Secretary Mrs. Ruby Young and Mrs. Joanne Baker; the technical support from our computer center, made my graduate study hopeful and easier.

Special thanks to my dear parents. Without their endless support and love for me, I would never achieve my current progress. They are my spirit and emotion supporters.

Finally, great thanks to my dear husband Yousu Chen, for his love and patience. It was my great fortune to meet and marry him. I'm always feeling happy and hopeful with him. He's my shining sun, always warm and cheer me up, even in darkness.

INTEGRATED DATA MODELING IN HIGH -THROUGHPUT PROTEOMICS

Abstract

by Shuangshuang Jin, Ph.D.
Washington State University
December 2007

Chair: John H. Miller

The purpose of this research project is to investigate the work flow in high-throughput quantitative proteomics. After data collection on complex protein mixtures subjected to proteolysis, liquid chromatography (LC) and mass spectrometry (MS), a long data reduction procedure beings that involves protein identification, protein abundance estimation, biological interpretation of differentially abundant proteins. The data reduction procedure contains many steps that are the subject of ongoing research in bioinformatics. This thesis research addresses the following issues: (1) protein database redundancy, (2) peptides from proteolysis that are found in more than one database entry, (3) separation of biological effects on protein abundance from variance due to instrument and processing effects, (4) data mining to relate observed global changes in an organisms proteome to biological processes perturbed by treatments.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Dedication

This dissertation is dedicated to my husband and parents

who provided both emotional and financial support.

Chapter 1

Introduction

1.1 Background and significance

Proteomics research started in 1994 when the term "Proteome" was first defined as "PROTEins expressed by the genOME" [1]. It is a field using quantitative methods to study the function and/or changes of all expressed proteins in a given cell, tissue, or organism under a variety of conditions. Since proteins are the most functionally important molecules involved in essentially all biological processes, proteomics research has a great significance in deciphering the mechanisms of gene expression controls and characterizing biological processes in terms of disease processes and drug effects [2]. To date, most proteomic studies qualitatively and quantitatively compare proteomes in terms of protein abundance or concentration in normal (control) and disease states in cells and tissues. By comparing protein expressions in different conditions, for example, saying that the presence of a particular protein at a particular concentration deviates significantly from a normal range of values, valuable information are obtained to understand the underlying physiology and pathogenic mechanisms [3].

Currently global proteomics approaches are challenged by the complexity of the mammalian proteome, the extensive range of protein concentrations and inadequate methods to quantify concentration differences. Although high-throughput mass spectrometry drives the proteomics research by changing the focus from the analysis of selected isolated proteins to proteome-wide analyses [4], difficulties still exist as a result in large-scale data manipulation, information extraction and pathway analysis. From the time the experimental raw data becomes available, there is a long way to go before we can really evaluate the significant role each particular protein plays in the biological system. Manual data acquisition and calculation on the enormous dataset is not practical and subject to unacceptable error rate. Without reliable data processing and statistical analysis, protein abundance change can not be calculated quantitatively, no confidence tests can be applied, and the subsequent significance evaluation and pathway analysis are not valuable.

Therefore, designing a seamless data flow to go through the processes, seeking a strategy to maximize useful sample dataset, preparing reliable data for statistical analysis to facilitate altered protein concentration estimation, and devising tools to test the consistency of proteomics data and its biological meaning, all have great significance to proteomics research.

An integrated data model, which contains all these properties as well as high speed, accuracy and automation, can serve as a good prototype in high-throughput proteomics research. It can help proteomics researchers to perform their proteomics analysis in a semi-automatic fashion and provide reliably identified peptides and proteins information and statistical analyzed data to interpret biological processes or test

biological hypotheses, which is also very important in disease understanding and detection, diagnosis investigation, and drug development.

1.2 Protein profiling

Assessment of differential protein expression from the observed properties of detected peptides is the primary goal of proteomics. In recent years, high-throughput technologies for measuring the global expression of different components of the biological system, such as genomics, proteomics, and metabolomics, have made significant progress [5-7]. RNA profiling using microarrays is accepted as the state-of-the-art approach to investigating genome-wide changes in gene expression. However, the use of gene expression patterns is insufficient for understanding protein abundance, as additional post-transcriptional mechanisms, post-translational modifications and degradation, may influence the level of a protein present in a given cell or tissue. Protein profiling using high-throughput mass spectrometry (MS) as a complement to microarray analysis increases the likelihood that genome-wide data collection will lead to the discovery and characterization of important disease pathways [8]. A major challenge in using these new proteomics technologies is devising ways to extract the full meaning and implications of the data in a semi-automated fashion that facilitates an understanding of the underlying biology.

Global proteomics approaches based on MS reliably identify peptides and proteins; however, using these data to determine which proteins in a complex mixture are up- or down-regulated by a given treatment or disease is critical to biological

interpretation of the data and is still challenging. The most common method for quantifying protein changes is 2-dimensional electrophoresis coupled with MS or tandem MS analysis. Here the gels are stained, the intensity of the spot used to obtain abundance information and the MS to identify the proteins. While this method has been used since the early 1970s, it has the disadvantages of being difficult to automate and its limited detection range misses low abundance proteins.

Since it is generally accepted that the intensity of the MS signal for a peptide may vary due to ion suppression effects between simultaneously eluting peptides, isotopic labeling techniques have been implemented in conjunction with MS approaches. Ostensibly this approach provides greatest accuracy but has several disadvantages which include 1) expense, 2) complicated labeling chemistries that sometimes yield artifacts, and 3) difficulties in identifying isotopic pairs for relative quantification. Recently there have been several reports describing "label-free" methods for LC-MS that successfully identify proteins with altered relative abundances [9-11]. These reports indicate that a linear correlation exists between the amount of peptide and its peak area when the LC flow rate is low and small amounts of sample allow for optimum electrospary ionization (ESI) efficiencies.

Liquid chromatography (LC) is a chemical approach that can separate a wide range of compound mixtures, from small molecules to peptides and proteins. Proteomic samples are usually enzymatically digested by trypsin to cut the amino acids lysine (K) and arginine (R) so that proteins are cut into smaller charged amino acid chains -- peptides. Under high-pressure chromatography, these peptide pieces are dissolved in a solvent and then placed on a column head, which is a tube packed with LC separation

media. Different species elute through the column at a different speed according to their resolving power to the solvent and separation material. Weakly-interacting species elute out of the column early while strongly-interacting species elute out of the column slowly. As a result, different groups of identical species arrive at the end of the column at separate times [12]. The resolving power (or peak capacity), which is defined as "the number of peaks that can fit into the length of the separation" [7], is a characteristic of LC separation approach. It is determined by several factors such as the LC column efficiency, gradient elution speed, particle size, or surface structures [13]. For analyzing complex mixtures, high peak capacity separation is always significant because of the limited number of species that can be resolved and identified by MS at any given time [7].

Mass spectrometry (MS) coupled with LC separation technique is widely used for analyzing highly complex protein mixtures. The LC column feeds the MS continuously with separated charged peptides. MS measures the frequency and mass/charge ratio of the peptides and fragments them into ions by collision induced dissociation, where inert gas, such as helium or argon atoms, is commonly used to carry out the collision. Collisional dissociation breaks the weakest links in the peptide backbone, which are NH-CH, CH-CO, and CO-NH bones [14]. Only charged fragmental ions generated by the collision process can be monitored by MS. Measuring the molecular weights of series of ions, the corresponding amino acid fragmentation can be reasonably predicted, so that the peptide sequence can be determined. Fragmentation patterns may differ for each mass spectrometer. Ionization technology, fragmentation technology, parameters setting, detector and instrument software may all have influence on it [12].

While shotgun sequencing (LC-MS/MS) is currently the most popular technique for large-scale protein profiling, another identification technology of growing importance is FTICR-MS, the Fourier transform ion cyclotron resonance mass spectrometers. It is a good complement to LC-MS/MS. In the first stage, enzymatically digested protein samples are cleaved into smaller peptides, and these peptide mixtures are analyzed by LC-MS to establish a set of tentative peptide identifications called Potential Mass and Time (PMT) Tags database, which contains the identified peptides as well as their exact elution time and mass information.

This PMT database is imported by running the MS/MS data against database search software for peptide identification. There are a lot of software tools available to assign peptides to MS/MS spectra: Mascot [15], MS-Tag [16], SEQUEST [17], and Sonar [18]. These database searching methods vary from each other on the algorithm design and required parameters, but all have the same role, that is, to search a protein database with a single peptide MS spectrum. They start with constructing theoretical spectrum for all peptides in a protein database, and then compare the experimental and theoretical peptide spectra using an effective scoring scheme. When the measured peptide MS/MS spectrum is matched with the database peptides within some tolerance, the measured peptide is initially identified. Identified peptides can be further filtered based on peptide ion charge (singly-, doubly-, or triply-charged), raw correlation score of the top protein candidate (Xcorr), difference in correlation score between the top and second peptide candidates (DelCN), and the tryptic nature of peptides to reduce incorrect peptide assignments. A discriminant program can also be used to determine peptide confidence probabilities. The discriminant score takes advantage of elution time information and

tryptic cleavage information, which enhances the accuracy of peptide identification [19].

In the next stage, the protein samples can be analyzed in duplicate by high mass measurement accuracy using FTICR-MS, coupled with the same type of separation. Then candidate peptide masses are compared to masses measured experimentally by the FTICR-MS. If a match with a small mass error and reasonable elution time can be found between a tentatively identified peptide and an experimental observed peptide, then that sequence become part of an Accurate Mass and Time (AMT) Tag marker database used to perform differential expression studies without further tandem mass spectral analysis. This match process is often called peak matching.

LC-FTICR is a good validation of the peptides identified by LC-MS/MS. The inherently greater resolution of FTICR over MS/MS also makes it able to identify low-abundance proteins with better sensitivity in proteome mixtures, which is difficult for MS/MS to identify with confidence. In addition, this proteomic technology has several other advantages. It has a higher dynamic range, better proteome coverage and higher throughput for large-scale proteome studies under multivariate conditions [7].

LC-MS/MS and LC-FTICR-MS play a significant role in high-throughput quantitative proteomics and protein profiling. They identify a list of peptides inside the complex mixtures, which are requirements for further protein identifications.

Inferring protein identities based upon peptide assignments is challenging. One must group all assigned peptides according to their corresponding proteins in the database [20], namely, find the shortest list of proteins needed to explain the detected peptides and make the most likely association of peptides with the proteins they identify. Detected

peptides can be unique, derivable from one protein only, or degenerate, possibly originating from several proteins. The homologous and redundant entries in database and sequence similarities of proteins are the main reasons for the occurrences of peptide degeneracy. This is a difficulty for protein identification, as the sequence of a degenerate peptide is present in several different protein sequences, they cannot identify the presence of a protein uniquely. Another problem comes from the false identification of a peptide, especially when that peptide uniquely identifies a single protein in the mixture. Therefore, measuring the protein identification confidence is necessary. Peptides probabilities can be combined as a factor to perform this validation. A number of protein identification software tools that can facilitate this assigning and filtering process are CHOMPER [21], DTASelect [22], INTERACT [23], Qsocre [24], and Protein Prophet [20].

A typical protein-peptides association can be further investigated by arranging proteins in classes according to the unique/degenerate character of the peptides that identify them. Greatest confidence is placed in proteins identified by multiple unique peptides and this is reflected in a confidence score. Proteins that cannot be distinguished by the peptides observed in the mass spectra are grouped together. These proteins generally have very similar sequence due to redundancy in the database of proteins for the organism under investigation or because they are isoforms. Additional information, such as the mass of the protein, is required to determine which protein among those is actually present in the sample.

Comparing a protein's concentration under different treatment is the focus of many proteomics studies. The abundance of a protein in a sample from a treatment group

relative to its abundance in samples from controls can be estimated from the abundances of the peptides that identify the protein, provided they are observed in samples from both the treatment and controls. Peptides of different sequence have intrinsically different detection probabilities for many reasons, including their ability to be ionized, which can greatly affect the areas under their peaks in mass spectra. If a peptide is observed in both treatment and control samples, the ratio of peak areas is independent of these instrument effects. This ratio should be the same for all unique peptides that identify a protein; hence, a simple average of the abundance of unique peptides relative to controls is a good estimate of protein abundance relative to controls. The abundances observed for degenerate peptides could be due to contributions from more than one protein in a biological sample; consequently, they are less reliable than unique peptides for estimation of protein abundance. Usually they are eliminated from protein abundance estimate for this reason. In Chapter 5, we will present our method of utilizing degenerate peptides for protein abundance estimates without introducing this ambiguity while including more peptides data.

Identification and characterization of signaling proteins whose expression is up/down regulated is a crucial step of protein profiling. The abundance of proteins from treated samples relative to controls can be estimated as part of an analysis of variance to return the upper and lower bounds on relative protein abundance for a specified statistical confidence level. Proteins judged to have abundances in treated samples that are statistically different from their abundance in control samples can be ranked by their abundance relative to control. In seeking a biological interpretation of abundance changes, priority is given to proteins with the greatest differential abundance. Typically,

proteins which have statistically significant differential abundance greater than 2 or less than 0.5 could be the initial targets. Databases are searched for gene-expression and intracellular signaling pathways that contain these targets. As pathways are discovered by this set of targets, we can reinforce the discoveries by searching the pathways for proteins observed in the experiments with less pronounced differential abundance and less confident identification. From the pathways discovered, hypotheses can be formulated regarding the biological processes responsible for the observed up- and down-regulated proteins. These hypotheses are tested by more traditional molecular-biology experiments that focus on a defined set of proteins and use methods of detecting abundance change that are more sensitive than mass spectrometry.

Protein profiling opens a door for system-level biology. Details on how the signaling pathway is involved to help us investigate the underling biological mechanisms are introduced in the following section.

1.3 Signaling pathway

High-throughput technologies enable us to collect comprehensive datasets on complex protein mixtures, protein profiling contributes to the identification and analysis of differentially expressed proteins in the sample, and signaling pathway helps us to understand the interaction of these proteins and uncover the underlying phenomena of this biological network at a system-level, which is the main purpose of system biology approach.

Systems biology can be defined as “an approach to biology where organisms and biological processes should be analyzed and described in terms of their components and their interactions in a framework of mathematical models” [25]. These components (e.g., in our case the differentially expressed protein targets) by themselves are not sufficient to understand the complexity of the organism, but investigating the regulation network containing these components will be much helpful for us to explore their functions and the biochemical interactions between them [26].

Database search against the differentially expressed proteins for signaling pathways that contain them is the main tool to construct the signaling network and interpret the protein’s abundance changes. Two categories of statistical approaches are usually used to rank the up or down regulation properties of the list of differentially expressed proteins: over-representation approach (ORA) and functional class scoring approach (FCS) [27]. ORA compares the number of differentially expressed proteins with the number of proteins expected to be found just by chance. If there’s a substantial difference, it is said to be significant. The probability of observing the actual number of proteins just by chance, for example, the p-value, can be calculated by this kind of statistical model. Alternately, FCS considers the distribution of the pathway proteins in the entire list of proteins and performs an enrichment analysis, which ranks all proteins based on the correlation between their expression and phenotypes and then calculates a score to reflect that [27]. Both of them are currently widely used, but they have limitations in system level dependencies and interactions exploration as well as pathway level perturbations and modification identification because their functional category is analyzed independently without a unifying analysis. Pathway databases such as KEGG

help to perform the analysis in a more comprehensive and powerful level.

KEGG [28-31] (Kyoto Encyclopedia of Genes and Genomes) is a database resource for systematic analysis of cells or organisms to understand their higher order functions from genome information. KEGG consists of three databases: GENES database, PATHWAY database, and LIGAND database. GENES database collects all the completely or partially sequenced genes and proteins. PATHWAY database stores the higher order functional information of genomes in terms of the interacting network. LIGAND database collects the information for chemical compounds and enzymatic reactions in the cell. KEGG helps in computerized representation and utilization of functional data, which are contained in the networks of interacting molecules.

Interpreting proteomics data for its biological meaning is our main target for proteomics analysis. One of the tools to construct the biochemical signaling network from experimental data and analyze them by computational methods to understand their role in complex biological processes is MetaCore [32]. MetaCore is an interactive database derived from manually curated literature publications on proteins and small molecules of biological relevance in humans. It was developed for the purpose of exploring biological interpretations with the integration of functional, molecular, or clinical information, and visualizing cellular components as networks of signaling, regulatory and biochemical interactions.

MetaCore provides several graph-based tools to relate proteins altered in abundance to biological processes affected by treatments. Currently it has seven network building algorithms and numerous options to specialize their use. In addition, data

filtration prior to network building based on tissue type, fold-change threshold, sub-cellular localization, etc. allow exploration of multiple scenarios for interpretation of a data set. The analyze-networks feature is designed to fragment super-networks of the input proteins into sub-networks with statistical scores that rank them according to their saturation with objects from the input data list. This feature is often used on large data sets without predefined restrictions to maintain the greatest flexibility of possible connections between proteins in the input list. This mode of analysis often produces many sub-networks that are well-populated with differentially abundant proteins, statistically significant, and biologically relevant. In proteomics study, this tool displays both the proteins that we identified, called "targets", and the direction of protein abundance change, called "up- or down-regulation" on signaling networks associated with the control of biological function. Hence, the network becomes a directed graph with nodes that are proteins and edges that indicate how a biological process affects their abundance.

MetaCore uses argument based on enrichment statistics (kind of FCS) to suggest which processes within its database of cellular signaling networks is the most likely explanation for the proteomic data. Enrichment statistics are based solely on the number of targets found on the network. The larger the number of targets the higher the enrichment score and the more likely the signaling network is a valid interpretation of the proteomic data.

The consistency of up and down regulation of targets with activation and inhibition within a network is not included in scores based solely on number of targets on a network. We propose that augmenting enrichment with consistency tests will increase

our ability to discriminate between networks in the MetaCore database which are possible explanation for our proteomic observations.

We know that to gain biological insight from protein profiling, it is useful to employ a software package like MetaCore to locate proteins, which are significantly altered in abundance, on known genetic and cellular signaling pathways. But first of all, we need to find identifiers for the proteins that can be recognized as nodes in the database of network interactions. Identifiers for human proteins include Entres Gene, LocusLink, SwissProt, RefSeq and Unigene. Identifiers for mouse proteins are limited to Entres Gene, LocusLink and RefSeq. The same type of identifier must be used for all proteins in the input; hence it is desirable to choose the type of identifier that maximizes the number of proteins that can be mapped onto networks. It is clear that effort to increase the number of differentially-abundant proteins mapped to networks is needed in some cases. This might be accomplished by choosing a different type of identifier or an alternate identifier of the same type, if one exists. A resource like Uniprot [33] can be used to find multiple IDs for the same protein. Batch conversion of identifiers can also be achieved using DAVID Bioinformatic Resources 2006, National Institute of Allergy and Infectious Disease (NIAID), NIH [34].

1.4 Outline of this dissertation

In this chapter, we have introduced the knowledge of the two most important concepts of proteomics research: protein profiling and signaling pathway. In next chapter we give an overview of the previous work on proteomics analysis, which our research is

closely related. In Chapter 3, a dataflow and an integrated model developed by us to facilitate the proteomics data processing procedures are described. The data analyzing processes of two functionally different proteomics studies using this model are demonstrated with results and discussions in Chapter 4. A detailed discussion of our approach on degenerate peptides processing is presented in Chapter 5, and possible directions for further research and development of the model are outlined in Chapter 6.

Chapter 2

Existing Methods and Previous Work

Mass spectrometry-based proteomics analysis coupled with liquid chromatography-based separation, is one of the most promising technologies to probe a complex biological sample globally across multiple conditions at the protein level. It holds great promise as a discovery tool for diagnostic biomarkers. In Chapter 1 we have introduced some basic concepts and technologies that involved in proteomics analysis. In this chapter, we are going to selectively illustrate some existing complete workflow solutions for LC-MS-based quantitative proteomics analysis [35-38], and two previous proteomics studies which we conducted before without integrating our new dataflow and proteomics data analyzing model [39, 40].

2.1 Existing methods

2.1.1 Quantification by linearity of signal ions and molecular concentration

To quantify proteins and metabolites by liquid chromatography-mass spectrometry, Wang and colleagues reported a new method in 2003 [35]. Without isotopic labeling or spiked standards, this biotechnology relies on linearity of signal

versus molecular concentration and reproducibility of sample processing, provides differential expression measurements and quantitative profiling of large-scale proteins and small molecules, and enables the discovery of biomarkers for clinical studies.

Wang et al. undertook a broad study to verify that analyte ion signals from electrospary ionization can in general reflect concentrations in a linear way even in the case of complex matrixes. They did observe the linear behavior in the signal from digested peptides of the synthetic mixtures corresponding to the protein concentrations. This linearity forms the basis of their analytical method for quantifying proteome and metabolome profile data for differential expression.

The quantification method they use relies on the changes in analyte signals directly reflecting their concentrations in one sample relative to another. It is based on the linearity as well as the stability and reproducibility of analyte signals. Spectral intensity normalization is employed during the quantification process to account for any long-term drifts in overall LC-MS response by employing signals of molecules that do not change concentration from sample to sample. An unbiased normalization procedure they use is based on determining the median of ratios of peak intensities of molecular components in the test sample relative to a reference sample and applying that median ratio as the normalization factor for each sample. A software application MassView was developed to perform this normalization to determine the constant intensity ratio between those unchanging analytes for the purpose of identifying the nonchanging concentrations.

This work established a direct quantification method to quantify proteomic and metabolomic profile data by LC-MS electrospary ionization without the need for isotopic labeling or spiking of special chemicals. It offers advantages such as simple sample

processing, applicability to proteome and metabolome samples, and less spectral interferences. Nevertheless, one restriction of this method is that the sample processing and LC-MS platform must be highly reproducible, which requires appropriate standard operating and system maintenance procedures. Significant computational cost is unavoidable for increasing sample complexity. Another point is that a threshold must be set properly so that all the signals being tracked have substantial ion counts to allow for the capture of both high- and low-intensity molecular ions.

2.1.2 An informatics platform for global proteomic profiling

In 2004, Radulovic and colleagues developed an informatics platform to integrate algorithms, statistical methods, and computer applications together to facilitate large-scale LC-MS-based gel-free shotgun profiling of complex protein mixtures [36]. Based on principles like experimental repetition, pattern recognition, and mathematical algorithms, this platform is a more advanced generation, and allows for systematic global comparison and classification of complex tissue proteomic samples, which further speeds up the discovery of biologically relevant proteomic biomarkers.

According to standard practice, peptide mixtures are subjected to LC-MS to form mass spectra. Extracting quantitative information from LC-MS datasets is the main contribution of this method. Several algorithms are applied. To filter signals from LC-MS raw data, Radulovic et al. developed a robust, assumption-free, threshold-like data filtering algorithm to detect real differences in peak number and intensity. A peak detection algorithm called “contour detection algorithm” is developed to automate peak definition based on boundary detection and integration techniques. A peak alignment

algorithm is used to correct peak drift and distortion to provide careful examination of protein abundance across multiple samples for the purpose of accurate biomarker discovery.

The platform permits meaningful quantitative and qualitative comparisons of proteomic datasets to identify differential protein expression between samples. And large-scale pattern recognition and mining of proteomic datasets are also automated to facilitate sample classification, which is the clinically important end-goal of expression profiling. Another significant advantage of this informatics strategy is that it is established on existing applicable LC-MS procedures and broadly available techniques and instruments to derive reliable protein profiling data, thus very convenient to use. One constraint of this method is its computational time. The alignment algorithm can provide careful examination, but it is computationally intensive and scales with the square of the number of experiments. When the sample sizes are large, the computational time is particularly demanding.

2.1.3 Signal maps

The combined method of mass spectrometry coupled with liquid chromatography is rapidly emerging as a method of choice for large-scale proteomics analysis. Usually, the probing of a complex biological sample is performed at the protein level. By determining the identities, abundances, and post-translational states of the myriad of proteins under different circumstances across multiple samples, similarities and differences can be identified, and expression profiling can be obtained to test the hypotheses regarding the biological roles of proteins in health and disease [41].

In the work of Prakash et al. [37], they went one step further and performed the comparison of complex biological samples directly on the signal level to resolve the problem that protein level comparison encountered: protein lists generated from individual experiments typically cover only a small fraction of the total protein content, making global comparisons extremely limited.

This work starts with constructing signal maps that associates experimental signals in the raw MS data across multiple experiments. The rule is to map the signals from any given peptide in the experiment to the signals acquired from the same peptide in the other experiment. Once constructed, this signal maps can be used for a variety of purposes in the furthering processing.

A lot of algorithms are involved in the implementation of this method. A score function is used to reward corresponding peaks. Alignment algorithm based on the score function is introduced to relate peaks of a run with peaks of another run through a signal map by choosing the alignment such that similar spectra appear close to each other in the sequence in the alignment. Two strategies: global alignment using a set of globally best pairwise alignments and progressive multiple alignment are explored to analyze the similarities and differences between different runs. Feature recognition method to detect features in real and virtual runs is also described in the work.

This signal maps approach is expected to decrease inter- and intra-experiment biases and improve the signal-to-noise ratio, thus be highly sensitive at identifying even low intensity signals and applicable to increasing throughput, which beats the common mass spectrometer technologies in their limited capability of sensitivity, reproducibility, and undersampling.

2.1.4 A review of comparative proteomic profiling methods

High-throughput mass spectrometry based proteomics technologies have made significant progress in recent years. Lots of approaches have been developed to facilitate this protein identification and profiling process to achieve the measurement of global expression of different components in biological systems for diagnosis discovery purposes.

In 2005, Listgarten and Emili [38] provided an overview of key statistical and computational issues relevant to bottom-up shotgun global proteomic analysis, with an emphasis on methods that can to some extent provide an accurate and rigorous assessment of the proteins' quantitative changes in their relative abundance in a complex biological sample.

In this review, several key directions of expression proteomics research have been readdressed as following: "Which proteins and variant isoforms are expressed during the lifecycle of an organism? Which post-translational modifications occur in each of these proteins? How do these patterns differ in different cell types and tissues and under different developmental, physiological, and disease conditions? How can biologists make use of this information to better understand the molecular basis for fundamental biological processes as well as for monitoring the course of disease so as to improve clinical diagnosis and treatment? [42-44]". These questions illustrate the difficulties of proteomics analysis on large-scale complicate biological system, and show the great importance of developing comprehensive methodologies and technologies to implement the complex profiling procedures.

Listgarten and Emili discussed a series of important computational and statistical concepts that should be considered when performing comparative proteomic analyses, and illustrated several quantitative analyses approaches for information extraction LC-MS based shotgun profiling, including the two we have discussed in the above sections (Wang et al. [35] and Radulovic et al. [36]). Moreover, they also outlined a typical sequence of operations for LC-MS datasets processing. It starts from quantization of peptide m/z values, signal filtering and background subtraction, amplitude normalization, peak detection and quantification, data transformations and error models, alignment in time, to classification algorithms and final biomarker discovery. Listgarten and Emili classified these steps into three levels of processing: Low-level, mid-level, and high-level, each focusing on different processing steps. For example, low-level involves data matrix formation, signal filtering, noise minimizing, and peptide quantization; mid-level involves data normalization, alignment in time, peak detection, peak quantification, peak matching and error models to facilitate profile comparisons; while high-level focuses on sample classification significance testing, and biomarker discovery, etc. Most of the LC-MS data processing approaches are based on this fundamental outlines with possible reordering of the intermediate steps, and different merits and limitations.

The basic statistical idea for examination of profiling datasets is mentioned in this review as well. Typically, when a statistics method is applied, it usually generates a score reflecting how much a feature discriminates between two classes. The distribution of test scores can indicate the information like the feature is discriminative or not.

As a future prospects of proteomic profiling, Listgarten and Emili anticipate that existing and emerging statistical and computational techniques with rigorous and

systematic evaluation will help unleash the full biomedical potential of proteomic profiling. Simultaneous LC-MS data alignment and normalization, and systematically tackling of preprocessing, classification, and biomarker discovery in a unified framework will definitely benefit the comparative profiling to a great extent.

2.2 Previous work

In this section, we are going to introduce two proteomics studies we have conducted before using LC-MS based proteomics technologies. The purpose of reviewing this previous here is to compare them with our new integrated model. Although these two studies have their own characteristics and processing and analyzing strategies, they have some commonalities and the same ultimate objectives with the new studies we are going to introduce in the next chapter. One of them is to use proteomic approach to identify and characterize protein shedding, and another is to identify shed proteins through cross-species manipulations. By looking into what methods these studies used to perform proteomics analysis, we can see the improvements and benefits of implementing our integrated dataflow on protein profiling and data analyzing, which will be presented in a good detail in the following chapter.

2.2.1 A proteomic approach to characterize protein shedding

Shedding is a mechanism by which cells change the repertoire of membrane proteins. Protein shedding is of great biological significance since it regulates a lot of biological processes. MS-based proteomic methods are usually the best tool for protein discovery.

To study the feasibility of identifying chemically-induced shed proteins, Ahram and colleagues [39] developed an optimal approach to perform sample processing and shed proteins identification using MS. Trypsin-digested protein samples are analyzed by reversed-phase capillary liquid chromatography interfaced to an ion-trap mass spectrometer. The resulting peptide spectra are analyzed by SEQUEST using a modified version of the human.fasta protein database provided by NCBI (National Center for Biotechnology Information). SEQUEST assigns the amino acid sequence of detected peptides to proteins to determine protein identifications. Microsoft Access 2002 is used to combine SEQUEST results and filter the peptide identifications based on criteria like peptide ion charge, correlation score, and tryptic natures. Protein abundances are then estimated using a peptide count method (count the number of peptides and proteins) based on the relationship between the abundance of proteins and the number of peptides observed by MS analysis. During this step, a set of rules are followed to identify changes in protein abundances: proteins must associate with cell surfaces (shed protein); and each protein must be represented by at least two different peptides to minimize false identification; protein must be identified by multiple experiments with multiple replications in treatments, etc. Furthermore, protein abundance estimates are validated using immuno-detection methods. Protein abundance changes in different treatments are also estimated.

This approach shows how to identify membrane proteins shed into the media using large-scale proteomic methods. It is a fundamental example for utilizing MS based proteomic technologies to investigate biological mechanism. However, additional precision and higher confidence information issues need to be considered.

2.2.2 Identification of shed proteins from Chinese hamster ovary cells

Ahram's group conducted another work regarding the identification of radiation-induced shed proteins from Chinese hamster ovary (CHO) cells [40]. Similar to [39], trypsin digested protein samples are separated by reversed-phase capillary liquid chromatography to separate proteins into peptides and analyzed by MS/MS. The difference of this study is that samples are also evaluated by FTICR-MS, which is several orders of magnitude more sensitive than ion-trap mass spectrometer. Again, SEQUEST is used to determine protein identifications. A discriminant function developed by Strittmatter et al. [45] is used to interpreter MS/MS data to increase confidence in peptide identification. SEQUEST results are imported into Microsoft Access 2002 to filter peptide identifications according to the similar criteria as described in [39]. Protein abundance estimates are also performed using a similar strategy with the only difference that the abundances are more accurate since they are obtained from FTICR analysis.

One characteristic of this CHO study is that it searches against cross-species protein databases (mouse and human) for MS/MS spectra interpretation, because a hamster protein database is not available, mouse-human homologs make it a good substitute for a hamster database. The success of identifying shed proteins of CHO validates that the high rate of protein homologs between mouse and human proteomes allows for the use of protein databases of closely related species to obtain cross-species protein identifications.

This study takes advantage of the high sensitivity of FTICR-MS to increase the number of shed proteins identifications, MS data are searched against both the mouse and

human databases, and a confidence scoring method based on discriminant analysis is developed to increase the positive identifications. It is possible to further advance it to conduct large-scale proteomic studies in the future.

Chapter 3

An Integrated Model for Proteomics Analysis

Genome-wide high-throughput mass spectrometry-based proteomics has emerged as an important new source of data on biological systems. This technology yields global information about the proteins expressed by an organism; consequently, biological processes can be studied without a priory assumption about the proteins that are involved. A profile of up- and down-regulated proteins is obtained which can be used to discover the gene-expression and cellular signaling pathways that underlie the disease state and/or response to treatment being investigated at the functional molecular level, which is of great significance in the discovery of diagnosis biomarkers.

Although identifying protein expressions and associating them with specific disease is one of the most promising areas of proteomics research [46], it is still challenged by the complexity of the mammalian proteome and the extensive range of protein concentrations. Many data-manipulation steps are involved in obtaining results of this type from mass spectrometry. This gives the data acquisition and processing a critical role in proteomics studies for complex protein mixtures.

The work of this dissertation is focused on prototyping a data flow to extract the full meaning and implications of the proteomic data in a semi-automated fashion, which involves data mining, developing methods to deal with missing data and degenerate peptides, integrating a rigorous statistical model to identify proteins with significantly altered abundances, and using the protein profile to characterize important signaling pathways to help us reveal the unique aspects of biological systems.

The resultant method will provide a seamless workflow for systematically constructing data and plots for proteomics analysis through data selection, classification, profiling, and interpretation processes. It will also provide an effective strategy to involve more identified peptides data to create a larger dataset for better statistical analysis, and a consistency-checking function as a complement to MetaCore software on pathway analysis. The experimental datasets we use were obtained from two studies. The first one is a toxicology study of mouse lung tissue and the second is a comparison of the mitochondrial proteome in normal and genome unstable cell lines. Detailed description of these two studies will be illustrated in the next chapter.

In this chapter, we will start with an overview of the dataflow we designed for these proteomics studies, followed by the detailed strategies we derived to solve each specific proteomics data processing and analyzing issues. Computer intervention and manipulation are involved. Mathematical and statistical methodologies are integrated.

3.1 Dataflow design

Mass spectrometry-based profiling combined with computer-based data processing is the main goal of our dataflow design. Figure 3.1 illustrates the flow of

information in protein profiling based on genome-wide, high-throughput quantitative mass spectrometry.

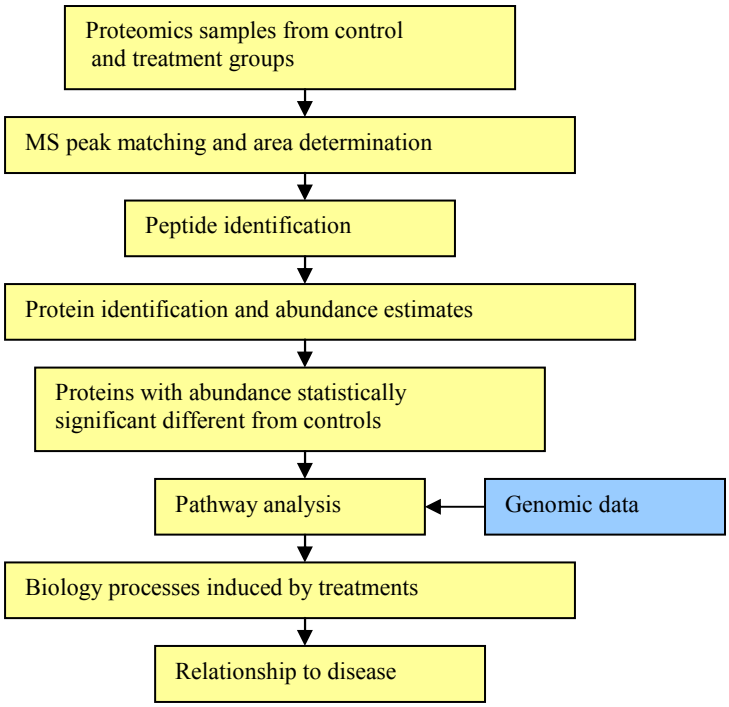

Figure 3.1: Information flow in protein profiling based on genome-wide high-throughput, quantitative mass spectrometry.

Biological samples from controls and treatment groups are subjected to enzymatic digestion to break proteins into peptides. Mass and fragmentation patterns are used to identify peptides and their relative abundance is determined from peak areas. Discriminate analysis combines observed mass and elution time with scores from database searching to yield an overall confidence score for peptide identification. Peptide identification is used to identify proteins most likely responsible for the detected peptides. Proteins are identified and the abundance of proteins from treated samples relative to controls is estimated under a specified statistical confidence level. Proteins judged to

have statistically significantly different abundance from the controls are studied in genomic pathways. Biological interpretations of the abundance changes are derived to achieve an understanding of the biological mechanisms and its relationship to diseases.

3.1.1 Peptide identification

Proteomic samples are digested and analyzed using tandem mass spectrometers coupled with high-pressure microcapillary liquid chromatographic separations. A composite of all samples (control plus treatments) are separated into fractions by strong cation exchange. Mass and fragmentation patterns from the MS/MS data are searched by SEQUEST v2.7 (ThermoFinnigan, San Jose, CA) against the National Center for Biotechnology Information (NCBI) protein database to identify the peptides present in the complex mixture. Parameters used in SEQUEST searches are variable, depending on the experimental design. The searches that generated data for this thesis were carried out using ± 3 Da restriction on parent mass accuracy and are unconstrained with respect to enzymatic cleavage, which allows for the detection of biologically modified peptides not normally associated with Trypsin digestion of proteins. (See Figure 3.2 for a sample SEQUEST parameter file.)

Sequest Parameters

```
[Sequest]
;DMS_Description =  --No Change--
first_database_name = C:\Database\M_Musculus_2005-12-08_NCBI.fasta
second_database_name =
peptide_mass_tolerance = 3.0000
create_output_files = 1
ion_series = 0 1 1 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
diff_search_options = 0.0000 C 0.0000 C 0.0000 C 0.0000 X 0.0000 X 0.0000 X
max_num_differential_AA_per_mod = 4
fragment_ion_tolerance = 0.0000
num_output_lines = 10
num_description_lines = 3
num_results = 500
show_fragment_ions = 0
print_duplicate_references = 1
enzyme_number = 0
```

```
nucleotide_reading_frame = 0
mass_type_parent = 0
mass_type_fragment = 1
remove_precursor_peak = 0
ion_cutoff_percentage = 0.0000
max_num_internal_cleavage_sites = 4
protein_mass_filter = 0 0
match_peak_count = 0
match_peak_allowed_error = 1
match_peak_tolerance = 1.0000
residues_in_upper_case = 1
partial_sequence =
sequence_header_filter =

add_Cterm_peptide = 0.0000
add_Cterm_protein = 0.0000
add_Nterm_peptide = 0.0000
add_Nterm_protein = 0.0000
add_G_Glycine = 0.0000
add_A_Alanine = 0.0000
add_S_Serine = 0.0000
add_P_Proline = 0.0000
add_V_Valine = 0.0000
add_T_Threonine = 0.0000
add_C_Cysteine = 0.0000
add_L_Leucine = 0.0000
add_I_Isoleucine = 0.0000
add_X_LorI = 0.0000
add_N_Asparagine = 0.0000
add_O_Ornithine = 0.0000
add_B_avg_NandD = 0.0000
add_D_Aspartic_Acid = 0.0000
add_Q_Glutamine = 0.0000
add_K_Lysine = 0.0000
add_Z_avg_QandE = 0.0000
add_E_Glutamic_Acid = 0.0000
add_M_Methionine = 0.0000
add_H_Histidine = 0.0000
add_F_Phenylalanine = 0.0000
add_R_Arginine = 0.0000
add_Y_Tyrosine = 0.0000
add_W_Tryptophan = 0.0000

[SEQUEST_ENZYME_INFO]
0.  No_Enzyme              0    -              -
1.  Trypsin                1    KR             -
2.  Trypsin_modified       1    KRLNH          -
3.  Chymotrypsin           1    FWYL           -
4.  Chymotrypsin__modified 1    FWY            -
5.  Clostripain            1    R              -
6.  Cyanogen_Bromide       1    M              -
7.  IodosoBenzoate         1    W              -
8.  Proline_Endopept       1    P              -
9.  Staph_Protease         1    E              -
10. Trypsin_K              1    K              P
11. Trypsin_R              1    R              P
12. GluC                   1    ED             -
13. LysC                   1    K              -
14. AspN                   0    D              -
15. Elastase               1    ALIV           P
16. Elastase/Tryp/Chymo    1    ALIVKRWFY      P
```

Figure 3.2: Sample SEQUEST parameter file. Parameter files may be customized to search for protein modifications if desired.

3.1.2 Preparation of PMT database

Results from SEQUEST searches are combined to give a set of tentative peptide identifications called Potential Mass and Time (PMT) Tags, which have been filtered based on peptide ion charge (singly-, doubly-, or triply-charged), raw correlation score of the top protein candidate (Xcorr), difference in correlation score between the top and second peptide candidates (DelCN), and the tryptic nature of peptides. A program developed by PNNL [9] is used to calculate a confidence score on peptide identification. The discriminant score takes advantage of elution time information in addition to SEQUEST scores, which enhances the accuracy of peptide identification. After building the PMT Database, distributions of discriminant scores and PMT quality scores are requested. These data are useful for determining a suitable cut off point when peak matching with the FTICR (Fourier transform ion cyclotron resonance) data to minimize false positive identifications. The PMT database functions as a lookup table of peptide-indexed (elution time, mass) pairs [47, 48] for later FTICR-LC-MS data (elution time, mass) comparison to reveal the underlying peptide's identity and its associated ion-current peak area [49-51].

3.1.3 Peak matching and LC-FTICR mass spectrometry

Because of the inherently greater resolution of FTICR over LC-MS/MS, the individual samples from control and treated animals are analyzed using a 9.4 Tesla LC-FTICR mass spectrometer. A separate Experiment is entered for each analysis. Data acquisition occurs only during the gradient phase of the LC run. The automated robotic ESI (electrospary ionization) interface allows for introduction of calibrant ions during the

last 5 minutes of data acquisition, thereby providing a run-to-run update of calibration coefficients and better mass accuracy. After the FTICR data are collected, mass to charge (m/z) ratios are extracted from the raw data using software developed at PNNL and listed in a single .pek file for a subsequent peak matching in the PMT database.

3.1.4 Q Rollup export

Q Rollup Export software [52] is used to filter the mass tag database and compile peptide and protein data. Only peptides having a Discriminant Score > 0.6 and a PMT Quality Score of 1 are considered. At the Q RollUp Export stage, these thresholds cannot be lower than they are at the peak matching stage.

Q Rollup speaks to the mass tag database using an SQL-based stored procedure, initiated through a GUI interface and user selection. Output is in the form of an Excel file, which is partitioned into several tabs. Prior to generating a Q Rollup file, the user may enter the "Edit/Define Q Rollups" tab and modify settings for each dataset. Protein tabs have a listing of all the protein database entries implicated by the SEQUEST searches along with the average abundance of all the detected peptide that identified each protein and some summary statistics. These protein abundance estimates are biased by instrument and processing effects that can be removed by the normalization procedure and statistical analysis discussed in section 3.1.6 and 3.1.7.

Peptide tabs have a listing of all the peptides detected in the samples by LC-FTICR-MS. There are redundancies because peptides are frequently found in more than one sample. Many peptides also map to more than one protein database entry name and all are included in this list. There are many scores associate with the peptide

identification; however, only discriminant scores are later used by Protein Prophet as a measure of the confidence of peptide identification.

3.1.5 Protein identification by Protein Prophet [53]

Digesting proteins into shorter peptides simplifies the MS/MS sequencing at the early stage of the process, but makes the assembling of peptide identifications back to the protein level a little difficult [54]. As shown in Figure 3.3, incorrect peptide identification (black squares) leads to false positive protein identification (black circles). Degenerate peptides may also be a reason for false positive protein identification. To solve this protein inference problem, i.e. the task of assembling the sequences of identified peptides to infer the protein content of the sample [54], we need automated database searching to help us determine the identities of the sample proteins. Protein Prophet [53] is one of the software tools we used to perform this task.

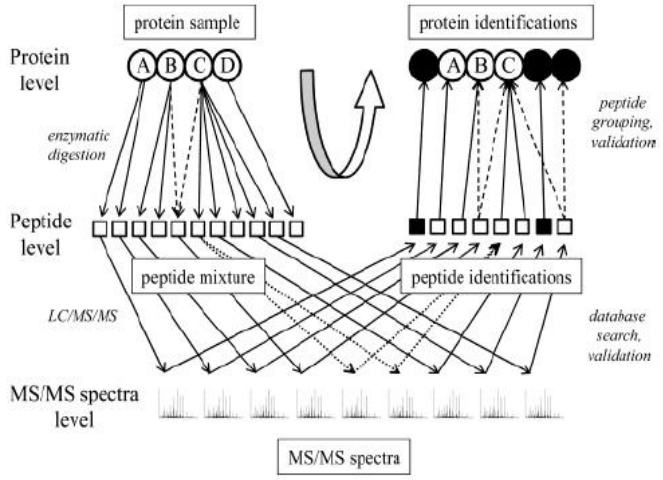

Figure 3.3: High-throughput mass spectrometry-based analysis of protein mixtures [55].

Peptide identifications gathered with Q Rollup are passed to Protein Prophet to find the shortest list of proteins needed to explain the detected peptides and to make the most likely association of degenerate peptides with the proteins they identify. Generation of a Protein Prophet input file is accomplished using Microsoft Access to query the Q Rollup output. Figure 3.4 shows the required fields for the Protein Prophet input file.

- $pep: peptide
- $cs: charge state used for abundance
- $XCorr: Xcorr score
- $Ref: Protein ID
- $pep2: the peptide formatted as: "-.pep.-"
- $degen: peptide degeneracy or the number of proteins that could contain the given identified peptide
- $prob: peptide prophet probability score applied to a given MS/MS hit or High MS/MS Descriminant Score
- $ntt: number of tryptic termini (0=no tryptic termini, 1= partially tryptic, 2=fully tryptic or 2 tryptic termini)

Figure 3.4: Required fields for a Protein Prophet input text file.

The fasta file of the protein database used by SEQUEST (NCBI database) is also a required input for Protein Prophet. Protein Prophet flags the unique peptides that identify a protein; thereby allowing proteins to be arranged in classes according to the unique/degenerate character of the peptides that identify them. Greatest confidence is placed in proteins identified by multiple unique peptides and this is reflected in a confidence score calculated by Protein Prophet.

Proteins that cannot be distinguished by the peptides observed in the mass spectra are grouped into the same "item". Proteins within an item generally have very similar sequence due to redundancy in the protein database or because they have similar biological function, such as different isoforms of a protein. Additional information, such

as protein masses, would be required to determine which proteins among those listed in an item are actually present in the sample. Since all proteins in an item are identified by the same set of peptides, abundance estimates derived from observed peptide abundances must be the same for all proteins in an item. Consequently, only one protein ID from an item needs to be associated with the peptide observations in creating the input file for statistical analysis to determine which proteins are significantly altered in abundance by the treatments.

3.1.6 Estimates of protein abundance

In addition to protein identification, quantitative MS allows us to estimate the abundance of proteins in a sample from observed peptide abundances. Other than treatment affects, the largest contribution to the variance of protein abundance comes from the intrinsic detection efficiency of peptides of different amino acid sequence. This contribution can be modeled as described below under the reasonable assumption that it is independent of biological variability and any treatment effects. When the variance due to peptide detection probability is removed by fitting the logarithm of peptide abundances to a linear mixed-effects statistical model, other instrument effects such as instrument drift and LC-column performance, are more easily seen in the data. They are described in Chapter 4 in the context of a particular data set.

A mixed effects linear statistical model fit to peptide abundance data by restricted maximum likelihood estimation [56, 57] was developed by the PNNL statistics group to estimate and quantify treatment effects in proteomic data. The model has two main components: the first describes the design of the biological experiment (sample injections

within animals within treatments) while the second describes the design of the sample processing (LC columns and peptide mass tags across injections). The latter component accounts for fluctuation in MS abundances due to column differences and, most importantly, differences in peptide measurability. Restricted maximum likelihood estimation (REML) offers a viable solution to fitting linear statistical models to LC-FTICR proteomics data that inevitably includes numerous missing ion current measurements for a significant proportion of the peptides across the multiple samples. The analysis leverages the many linear modeling tools immediately available to fit models, estimates errors and confidence intervals, diagnosis the model fit to the data set, and presents results in established and well understood formats.

3.1.7 Statistical analysis of protein relative concentration estimates

The LC-MS measurability of distinctly sequenced peptides is intrinsically different for many reasons, including differential digestibilities, elutions, and ionization potentials. Nevertheless, the relative difference in ion current between one peptide and another of equal molarity because they are unique to the same protein is consistent across samples. This observation suggests a multiplicative statistical model for a peptide MS abundance measurement: A = CPE where A is measured abundance, C is the peptide's concentration under the treatment, P is the effect of peptide measurability and E is random measurement error. It follows that if the abundance of peptide P is measured in both treatment (At) and control (Ac), the expected value of its abundance ratio is the peptide's relative concentration: At/Ac = CtP/ CcP = Ct/ Cc. This ratio should be the same for all unique peptides that identify a protein because the unique peptides of a

protein within a sample should be at equal molarity. Hence, for a given protein, a simple average of the ratios of its unique peptides under a treatment relative to the control is a good estimate of the treatment's protein concentration relative to the control. This simple model is the basis for the more complex statistical treatment developed at PNNL using a REML-fit mixed-effects linear model of log-transformed abundances that includes terms reflecting both the design of the biological experiment and the LC-MS sample processing.

A statistically rigorous approach based on a mixed effects linear model [58] was used to assess significant protein abundance change by the treatments. For this analysis, the following input files are required: (1) the Pedigree File, which describes the experimental design of the data, and (2) the MS File, which is generated by combining information from Protein Prophet output file with the peptide crosstab from the Q Rollup file. After statistical analysis, identified proteins are divided into following groups: (1) proteins for which there are insufficient peptide data to accurately determine the parameters of the statistical model, (2) proteins for which an abundance in control samples can not be estimated due to insufficient peptide data, (3) proteins with abundances in one or more treated samples that is significantly different from their abundance in control samples, and (4) proteins whose abundance is not significantly changed by any treatment. The 2+2 rule (at least 2 observations of a peptide in at least 2 groups, control or treated) is used to define group #1 proteins. Since abundance relative to control cannot be calculated for members of group #2, a different statistical model [59] is used to determine if a member of this group was significantly up regulated by one or more treatments.

When an experiment involves more than one type of treatment, the abundance of a protein may be significantly modified by more than one treatment. To eliminate this redundancy, proteins with altered abundance are sorted into the unique regions of a Venn diagram. For a 3-treatment study, the Venn diagram has the 7 regions illustrated in Figure 3.5.

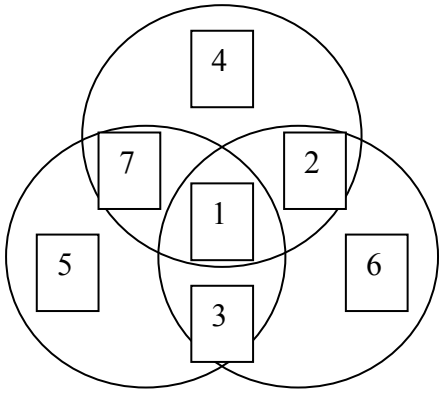

Figure 3.5: Venn diagram showing the overlap of proteins with altered abundance in three treatment groups.

3.1.8 Biological pathway analysis

The software package MetaCore™ (version 3.2.1 Copyright © 2000-2006 GeneGo Inc) is used to find biological interpretations of groups of proteins altered in abundance by the various treatments. Inputs to MetaCore are prepared by associating the logarithm of abundance relative to control with 2 gene identifiers, RefSeq and Gene Symbol, for all proteins judged by the statistical model to be significantly up- or down-regulated. Inputting the logarithm of relative abundance allows MetaCore to display the directionality of abundance change on signaling networks that involve observed proteins. Providing more than one identifier for each protein in the input list increases the

probability that MetaCore recognizes the protein as part of its database of signaling networks derived from ongoing literature surveys.

3.1.9 Summary of the dataflow

The designed dataflow discussed above facilitates semi-automated proteomic data analysis to provide a reliable dataset of up- and down- regulated proteins for biological pathway analysis. Several issues remain to be explored inside the data model, which will be discussed in Chapter 6. In next chapter we'll illustrate the use of the current data flow to proteomic data sets.

3.2 Software implementation

High-throughput MS projects lead to a large amounts of data that need to be manipulated and analyzed. To cope with the need for automated data conversion, classification, and filtering in the field of proteomics analysis, we developed an software tool ProteoViz. The system handles data for each individual steps, automates data creation, and provides a Java graphical user interface for managing the data manipulation steps which leads to the pathway analysis data from the initial identified peptides data. Perl scripts are called by the Java GUI to perform data and information extraction. Furthermore, bar charts are created automatically to enable proteomics researchers to better understand the data and interpret the underlying biological meanings. The work flow of ProteoViz is outlined in Figure 3.6.

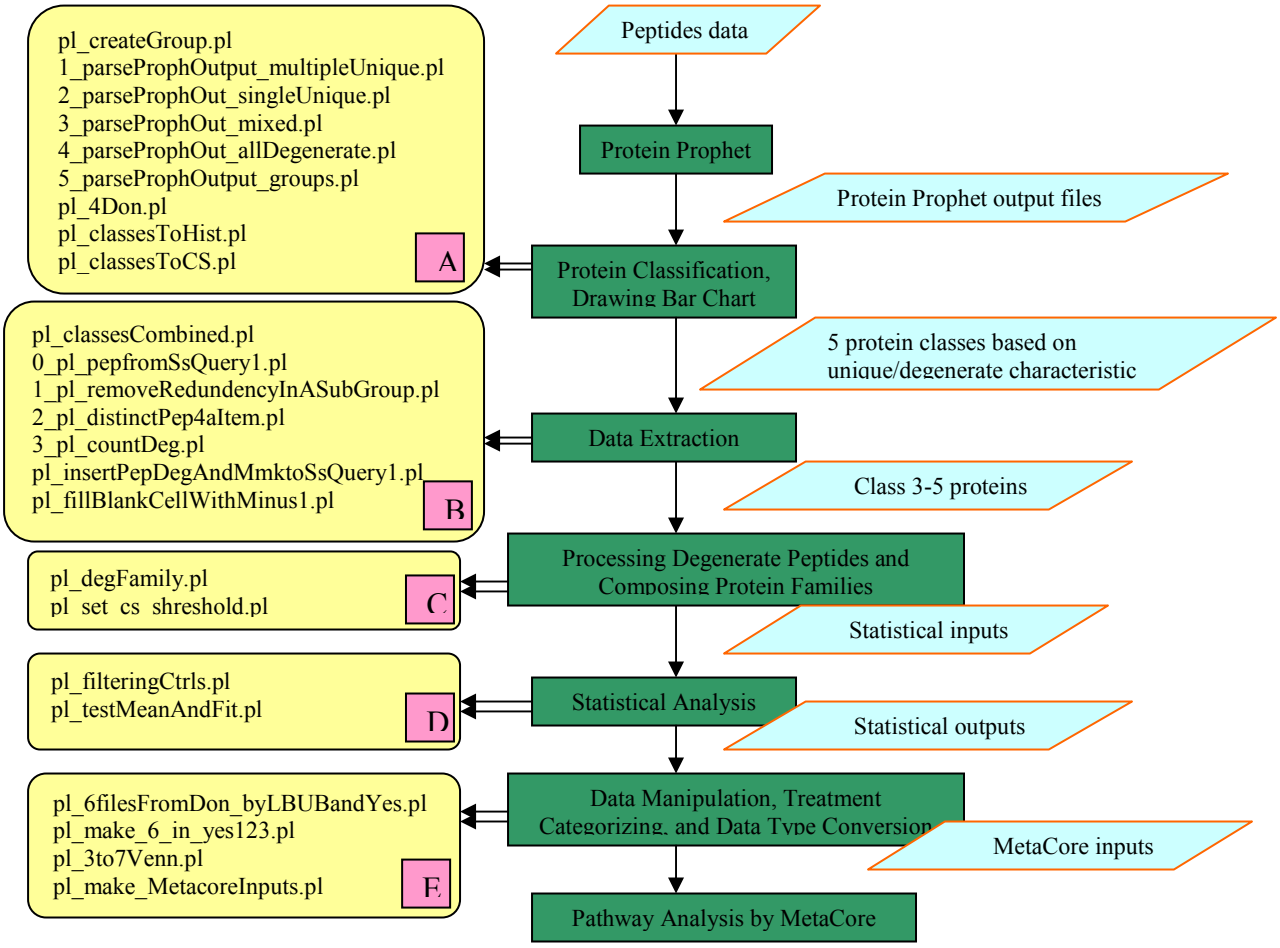

Figure 3.6: Work flow in of ProteoViz. Green rectangle indicates running process. Cyan parallelogram indicates input/output operation. Yellow oval indicates specific executing operations involved in a particular process. Magenta squares are block indexes for the processes.

3.2.1 Running Protein Prophet and parsing the data to classes

Data on identified peptides from Q Rollup are arranged in a required format as an input to Protein Prophet for protein identification. We then differentiate the Protein

Prophet identified proteins in 5 classes according to the unique/degenerate character of the peptides that identify them:

1). Protein identified by multiple unique peptides,

2). Protein identified by single unique peptides,

3). Protein identified by mixed peptides,

4). Protein identified entirely by degenerate peptides,

5). Protein identified by groups.

In each class, protein prophet confidence score, item number, and protein reference ID are associated. Histogram and confidence score bar can be plotted based on the classes' data in this step (Figure 3.6, Block A).

3.2.2 Wrapping all classes of proteins into full protein-peptides abundance dataset

Then we combine these protein classes into a big dataset, attach peptide information for each protein entry, remove protein and peptide redundancies, count degeneracy, and insert degeneracy number and molecular weight. These compose an initial input dataset for statistical analysis (Figure 3.6, Block B).

3.2.3 Applying the degeneracy approach and running statistical analysis

We use Perl scripts to compose protein family closures based on class 3-5 proteins in the initial input dataset and decrease protein degeneracy according to our degeneracy approach presented in Chapter 5. Statistical test and analysis are applied to identify proteins with significantly altered abundance. Figures are plotted to facilitate data analyzing (Figure 3.6, Block C and D).

3.2.4 Analyzing statistical analyzed dataset and Make MetaCore input for pathway analysis

Proteins with statistically significant different abundance in treated samples relative to controls are studied for their biological functions. They are sorted into 7 unique regions of a Venn diagram (Figure 3.5) and loaded for pathway analysis separately to allow for a better interpretation of the biological processes affected by each individual treatment (Figure 3.6 Block E).

3.2.5 The goodness of ProteoViz

ProteoViz integrates the data processing steps together, enables serialized data manipulation to better utilize the original experimental data. More useful information is extracted. Meaningful expressions of data are achieved. They facilitate semi-automated proteomics analysis to provide a reliable dataset for biological pathway analysis and disease findings.

The goodness of this tool can be characterized as following:

1). Comprehensive: ProteoViz provides the necessary functionality for our proteomics analysis dataflow. Protein Prophet is combined to make protein-peptides association. Statistical methods are integrated to enable data analyzing. Histogram and bars are drawn to visualize the data. And required dataset in handy format are created automatically for further pathway analysis.

2) Simple: Tedious data manipulation steps are serialized in ProteoViz. After loading the source file, ProteoViz can follow the dataflow to automate the creation of

required dataset. People don't have to create several Access databases to perform all kinds of operations on data for an expected dataset.

3) Reliable & Fast: Instead of manually data preparation, which is slow and prone to human error, ProteoViz produces more reliable results in far time. For a dataset with 3000 proteins, Proteoviz can create the MetaCore source file in three hours (Most of the time are occupied by running the Protein Prophet) while an experienced data analyzer may take a whole day or more to accomplish. With this quality, we believe Proteoviz is a good assistant for proteomics researchers, even those who is not familiar with computer operations. In figure 3.7-3.9, several screen shots of ProteoViz are given to illustrate the basic functionalities of this tool.

3.3 Summary

In summary, we have prototyped an integrated data model to fulfill the high-throughput proteomics analysis. Data manipulation, information extraction, missing data and degenerate problem exploration, and statistical model devising are all discussed in a good detail. We hope this semi-automated data flow will facilitate the proteomics research by computer modeling, fully extract the underling meaning of proteomics data, and unravel biological processes in favor of disease investigation and development of new-methods of treatments.

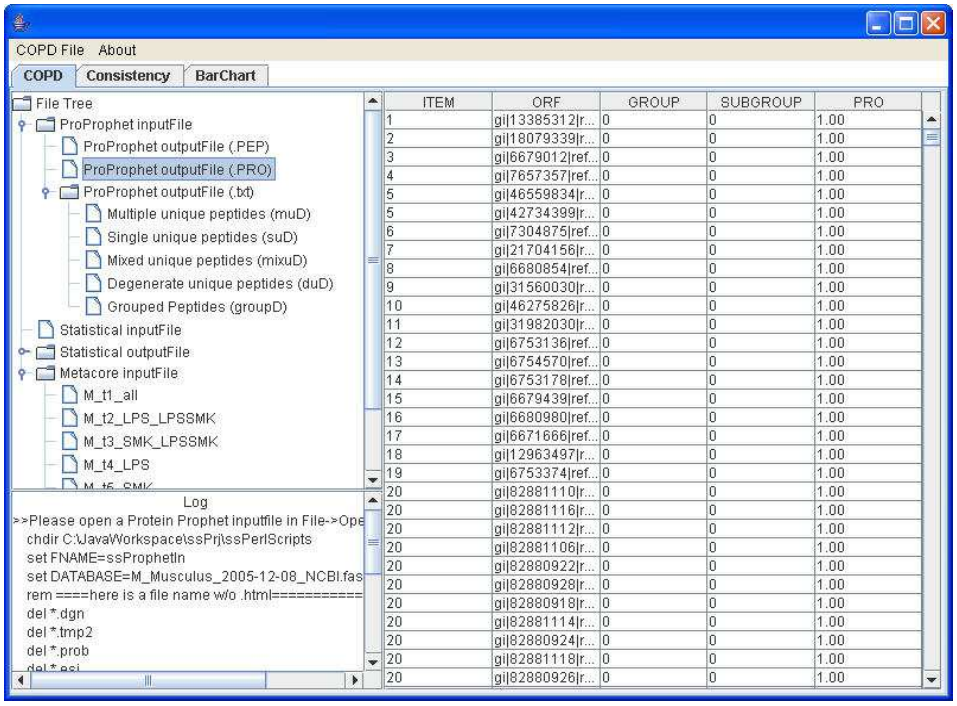

Figure 3.7: Data manipulation by ProteoViz, including file tree, log of data flow execution, and data visualization window.

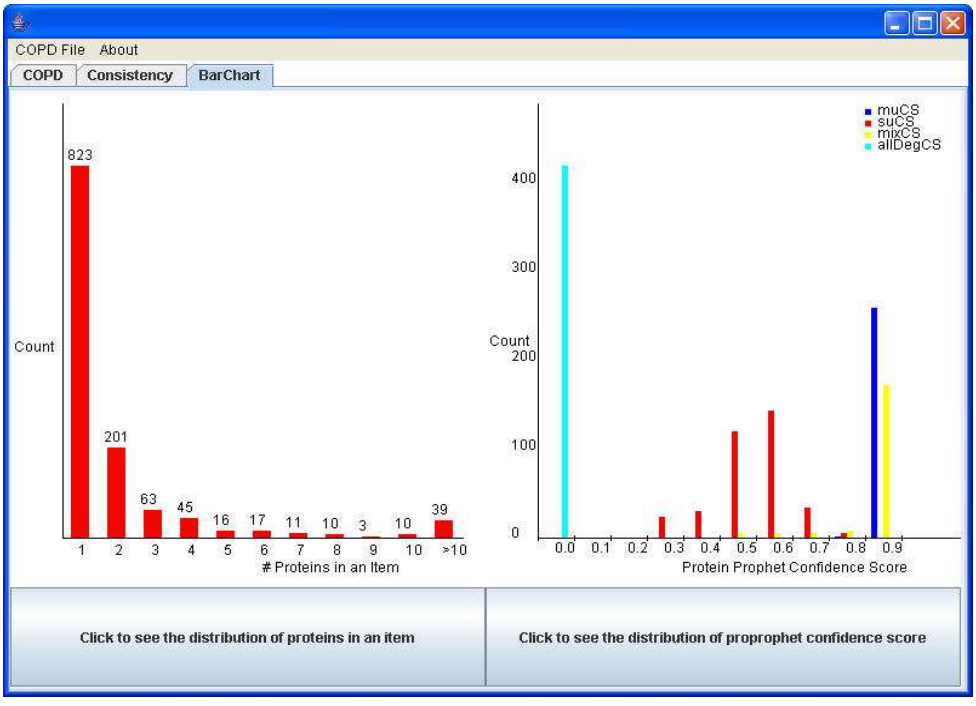

Figure 3.8: Histograms created by ProteoViz, including distribution of proteins in an item, and distribution of Protein Prophet confidence score.

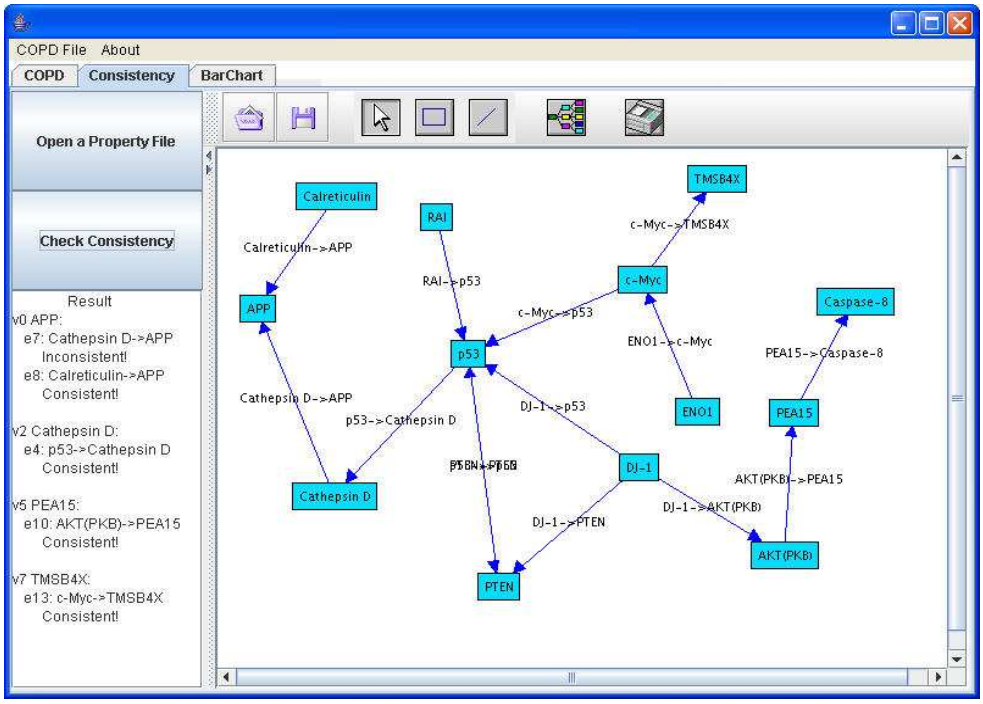

Figure 3.9: Reconstructed graph in ProteoViz.

Chapter 4

Applying the Integrated Dataflow Model for Proteomics Studies

Genome-wide high-throughput mass spectrometry-based proteomic technologies have made significant progress in protein identification, profiling, and measurement of global expression of different components in biological systems. We have demonstrated in Chapter 3 our seamless data flow to extract the full meaning and implications of the data in a semi-automated fashion. In this chapter, we illustrate the characteristics and benefits of using this data processing and analyzing method in proteomics research work. The experimental datasets we use are obtained from two studies. The first one is a toxicology mouse study relative to chronic obstructive pulmonary disease (COPD) [60-62]. Another one investigated mitochondrial proteome in cells exposed to radiation-induced genome instability (RIGI).

4.1 COPD study

The purpose of this study is to investigate whether the co-exposure to cigarette smoke (CS) and inflammatory-inducing lipopolysaccharide (LPS) will heighten the pulmonary lesions [63-65] in mice, and thus create a mouse model as a surrogate to mimic COPD in human smokers [66-75].

Development of COPD in rodents treated with CS alone takes long exposures and results in mild lung lesions, which limits the usefulness of the animal model for mechanistic research and therapeutic development [76]. Evaluations of genomic, proteomic, and classical toxicological end points in the early stages of pathological changes in the respiratory tract of mice exposed to CS, LPS and combined exposure were carried out to after a short exposure to determine the exposure regiment for subsequent chronic COPD experiments.

The proteomic data collected in the integrated mouse lung-tissue study provided a large complex dataset to investigate label-free methods to identify proteins with statistically significant abundance changes induced by the 3 exposure regiments.

4.1.1 Materials and methods

4.1.1.1 Study design

The detailed experimental design and results for clinical chemistry and histopathology were reported by Lee et al [77, 78]. Briefly, thirteen week old male AKR/J mice were exposed via nose-only inhalation for three consecutive weeks via one of the following regimens: 1) LPS (Lipopolysaccharide), 2) CS (Cigarette smoke), 3)

CS/LPS (Smoke plus LPS). Sham controls were exposed to high-efficiency particulate air (HEPA)-filtered humidified air.

4.1.1.2 Lung sample preparation for proteomics

Lung samples were collected at the end of three week exposure and were weighed before being tied for division. The left lung lobe was dedicated to our proteomic analysis. The proteomic samples contained lung tissue sampled from 5 mice per group (LPS, CS, CS/LPS and controls). Replicate injections for each of the 20 samples (one sample was run in triplicate) were analyzed by the LC-MS/MS; thus a total of 41 MS runs were evaluated. The samples were queued systematically for LC/MS-MS analysis, ordered by treatment and then animal within treatment. The two samples from each animal were queued sequentially to the two LC columns. Sample processing, which included two 24-hour periods for routine maintenance, required 7-days.

4.1.1.3 LC/MS-MS & LC-FTICR analysis and peptide & protein identification

LC/MS-MS & LC-FTICR analysis and peptide & protein identification were performed according to the dataflow discussed in Chapter 3. Briefly, analysis of enzymatically digested protein samples using tandem mass spectrometers coupled with high-pressure microcapillary liquid chromatographic separations was performed to establish a set of tentative peptide identifications called Potential Mass and Time (PMT) Tags, next the individual animal samples were analyzed in duplicate by high mass-measurement-accuracy Fourier transform ion cyclotron resonance mass spectrometers (FTICR-MS), coupled with the same type of separation system, so that peptide elution

time information can be used. Candidate peptides in the PMT were compared to masses measured experimentally by the FTICR-MS. If a match with a small mass error and reasonable elution time was found, then that peptide became part of an Accurate Mass and Time (AMT) Tag marker database used to perform differential protein expression studies without further tandem mass spectral analysis. Protein Prophet made the most likely association of the peptides with the proteins they identified. ProteoViz classified these proteins with respect to the unique/degenerate properties of the peptides that identified them, manipulated the data into specified formats, and applied statistical analysis to the protein abundance data to evaluate the up/down regulation characteristics of the proteins under different treatments. Pathway analysis upon the proteins with the most significantly altered abundances was performed in MetaCore to investigate the underlying mechanisms of the biological system.

4.1.2 Results and discussion

4.1.2.1 Peptide and protein identification

The distribution of discriminant scores for peptides in the PMT database (Figure 4.1A) has a large peak at very low scores. Including these peptides in peak matching of FTICR mass spectra would likely result in a large number of false positive identifications from purely random hits [79]. Excluding this peak, the distribution of discriminant scores is relatively flat until a second broad peak emerges with a maximum near 0.8 (Figure 4.1B) which identifies the most reliable portion of the original PMT database assembled by SEQUEST analysis of LC-MS/MS spectra. Consequently, a smaller PMT database

was assembled from mass tags (peptides) with discriminant scores greater than 0.6. This reference PMT database contained about 3800 peptides.

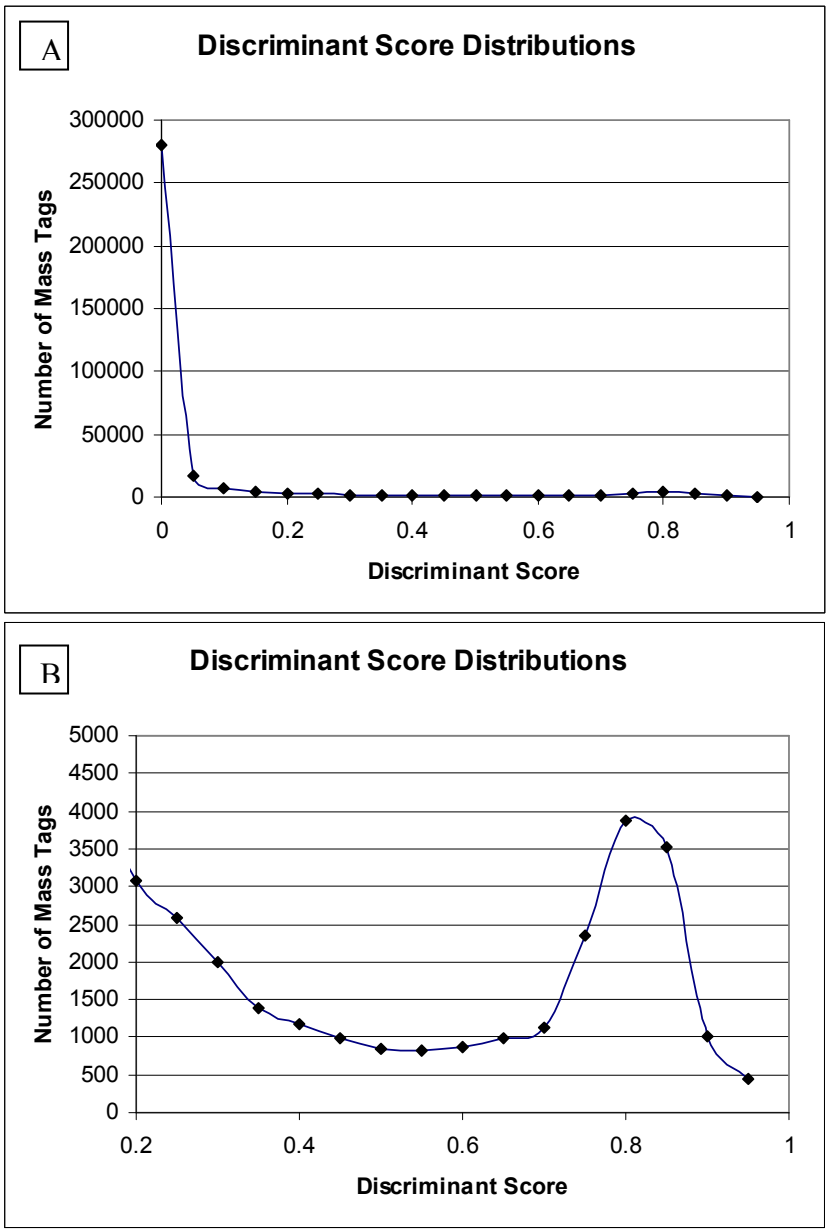

Figure 4.1: Distribution of discriminant scores among peptides identified by SEQUEST analysis of LC-MS/MS spectra using biological samples from control and treated mice. Panel A represents all peptides identified by SEQUEST and panel B shows only those peptides with discriminant scores ≥ 0.2.

The mass and time tag pairs from the FTICR-MS analyses of the 41 individual samples were matched against the reference PMT database, identifying 3219 candidate peptides, each appearing at least once across the 41 LC-MS injections. These peptides mapped to 2834 mouse proteins in the NCBI database, which Protein Prophet separated into 825 unambiguously identified proteins plus 415 groups of proteins which have similar amino acid sequence and could not be distinguished by the detected peptides. Figure 4.2 shows that about half of the 415 groups are pairs of proteins, which in many cases consisted of a known protein and a theoretical homolog. Many of the larger groups are composed of different isoforms of a protein. Since all members of a group of indistinguishable proteins are identified by the same set of peptides, abundance estimates derived from the mixed-effects statistical model apply to all members. Hereafter, we will refer to all 1240 items in the Protein Prophet output (825 unambiguously identified proteins and 415 groups) as simply "proteins".

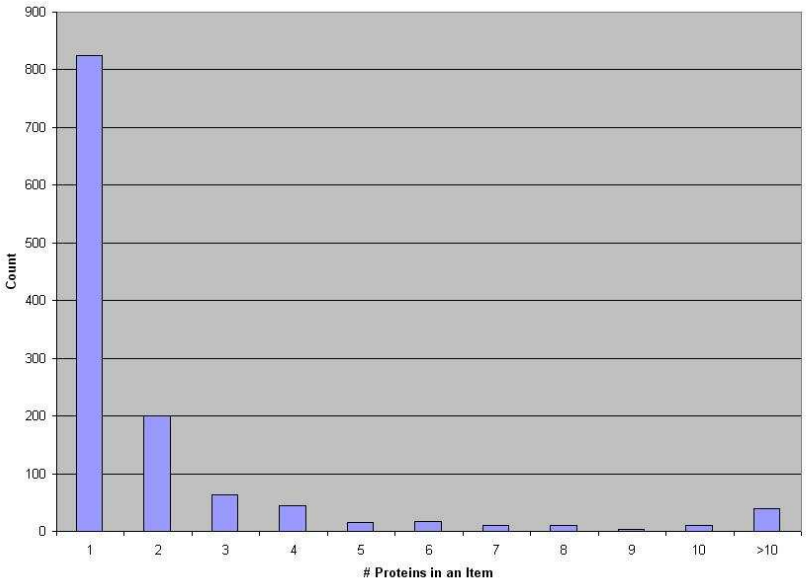

Figure 4.2: The distribution of the number of proteins in items of the Protein Prophet output.

A peptide was called “unique” if it identified a single protein or only one of the 415 groups of indistinguishable proteins with highly similar sequences due to a close biological relationship (i.e. homologs and isoforms). About one third of the 3219 peptides detected in the LC-FTICR spectra did not pass this test and were called “degenerate”. The 1240 proteins in the Protein Prophet output were grouped according to the unique/degenerate feature of the peptides that identified them. Figure 4.3 shows the distributions of Protein Prophet confidence scores in these groups. As expected, greatest confidence was assigned to proteins identified by multiple unique peptides (Class 1). Proteins identified by a single unique peptide (Class 2) usually have a lower confidence scores than those identified by a mixture of unique and degenerate peptides (Class 3). Proteins identified by degenerate peptides only (Class 4) were usually assigned very low confidence scores.

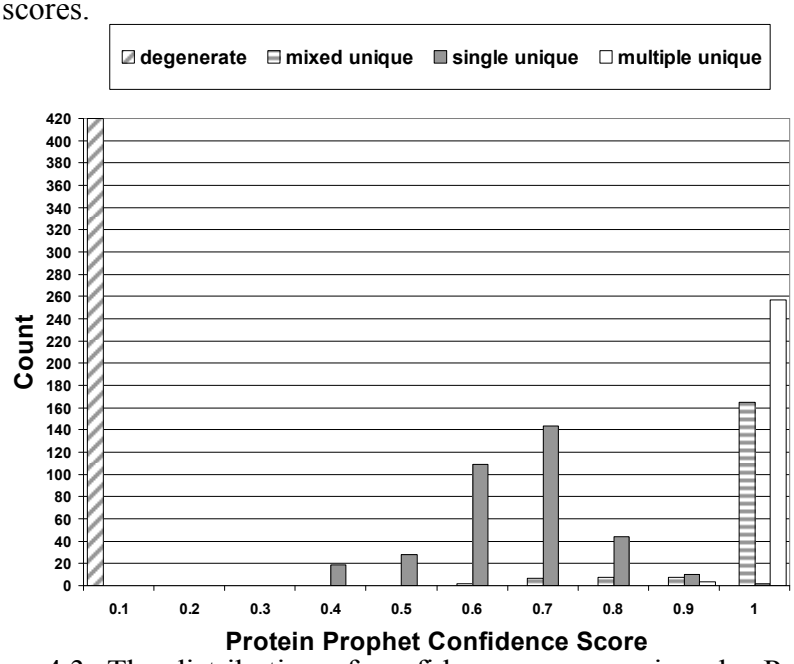

Figure 4.3: The distribution of confidence scores assigns by Protein Prophet to proteins identified by multiple unique peptides (open), single unique peptides (closed), and a mixture of unique and degenerate peptides (stripped).

4.1.2.2 Instrument effects

The COPD study contained 20 lung-tissue samples from 5 control animals and 15 treatment animals, 5 in each of the 3 groups: LPS, CS, and combined CS/LPS. Two sub-samples were drawn from each sample and injected into the LC-MS/MS. A third injection was made for one sample which meant that a peptide could be detected in as many as 41 MS runs. However, many of the detected peptides were not seen in all 41 runs due to either low abundance or misidentification. The statistical methods used in this work were specifically designed to treat unbalanced data sets that result from missing data.

Ion current measurements of peptides present at equimolar concentrations may vary significantly [80] due to differences in LC-MS measurability. Some peptides are cleaved more consistently than others, some peptides elute better, and some ionize easier. Differences in peptide measurability are the largest source of variability in mass tag ion current measurements. Since peptide LC-MS measurability is independent of biological variability and treatment effects, it can be accounted for by a mixed effects linear statistical model [81] as discussed in Chapter 3.

When the estimated effects of peptide measurability are removed from the log-transformed ion current measurements, other instrument effects are more easily seen. Two other processing effects are immediately apparent in Figure 4.4A. First, treatment groups were run in sequence beginning with controls. The 6 samples from control animals that were run on Sunday have median abundances that are below the overall median, indicated by the heavy horizontal line, while 3 of the remaining 4 control

samples run after a 24-hour routine maintenance period have median abundances above the overall median. In retrospect, the impact of these day-to-day processing effects could have been reduced by a blocking technique that ran samples from controls and treated animals in groups rather than running all controls followed by all LPS-treated animals, etc.

The second processing effect revealed in Figure 4.4A is a dependence of peptide abundances on the LC column, which is most evident in data from the middle period of acquisition where alternating light and dark shades reveal a systematic difference from alternating columns on the dual-column instrument. Alternating the two LC columns in sample processing is, in effect, blocking; hence the column effects could be directly accounted by our linear mixed effects statistical model [81].

To reduce the day-to-day processing effects that could not be included in the statistical model, the measurements were normalized by a procedure that is commonly used in the analysis of microarray data (Figure 4.4B). The procedure is more problematic for protein profiles due to missing data; nevertheless, the short exposures in the mouse-lung COPD study that generated our dataset makes it reasonable to assume that many of the observed proteins were only weakly affected by the treatments and justifies normalization based on medians.

4.1.2.3 Relative concentration estimates

Of the 1240 proteins determined by Protein Prophet to be the minimum number necessary to explain the observed peptides, 303 lacked a sufficient number of peptide abundance observations (at least 2 peptides measured in samples from at least 2 animal

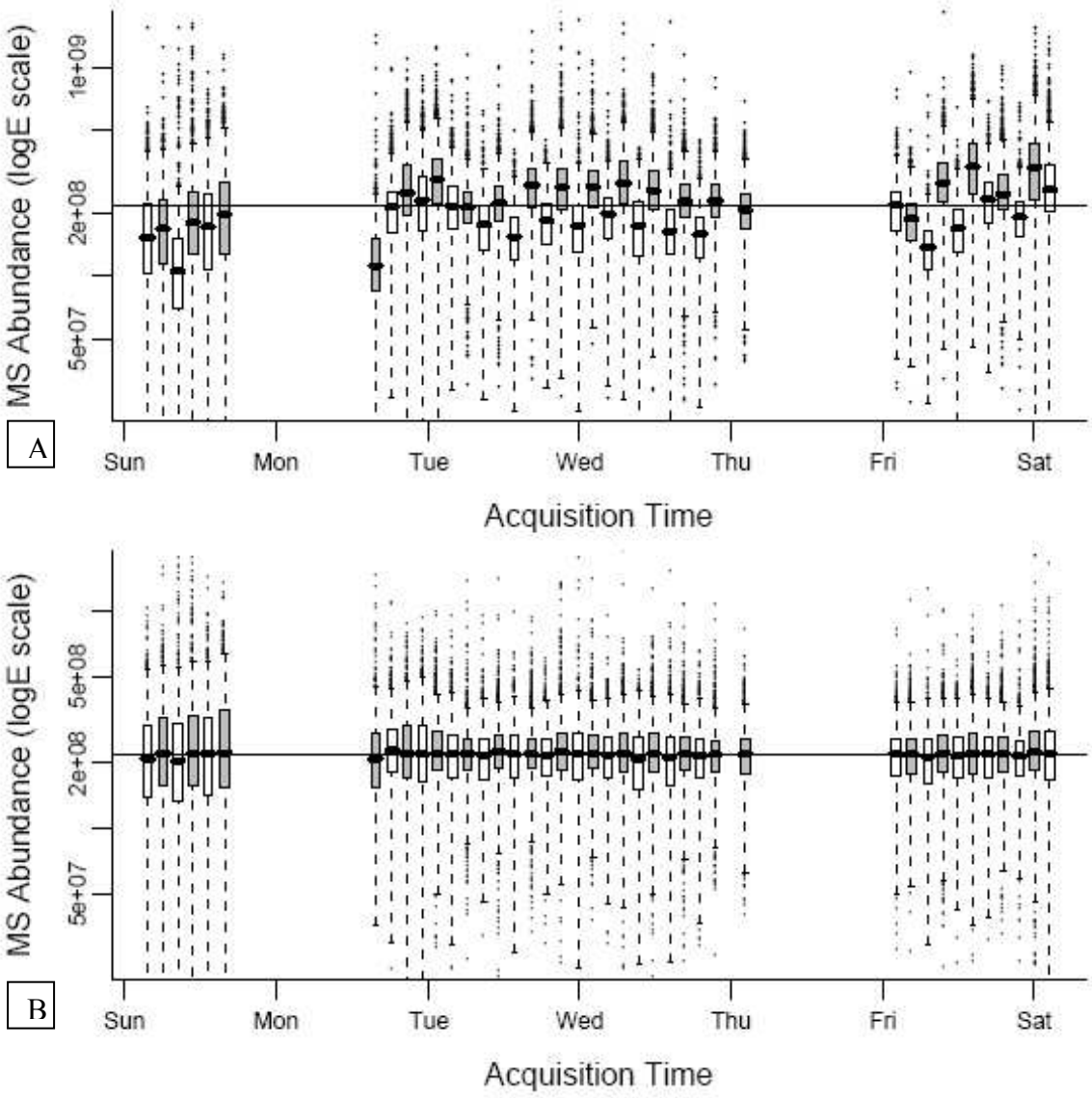

Figure 4.4: Logarithm of peptide abundances plotted as a function of injection times for FTICR-MS analysis. A box indicates the range of peptide abundance that contains 50% of the data with dashed lines denoting the range of approximately 90%. The sample median is marked by the bar across each box. Outliers are plotted as individual points. Light and dark shading indicates samples run on different columns. Panel A shows the raw data. Panel B shows the data after a normalization procedure based on medians.

groups; 2 + 2 rule), to estimate the parameters of the mixed-effects statistical model. Cases where numerous peptides were detected in samples from a single animal group, control or treated, would be rejected by the 2 + 2 rule; however, no cases of this type were present in the peptide data. In 78 of the 937 cases where the 2 + 2 rule was satisfied, calculations of relative protein concentrations in treatment groups were not possible due to a failure to detect sufficient peptides in samples from control animals. This type of observation suggests up-regulation of proteins by treatment effects; however, the observation may not be statistically significant. To test for statistical significance in these 78 cases, a generalized linear model [82] of measurement presence/absence was developed [83] to test the null hypothesis that the proportion of protein observations (number of observations of a peptide identifying a protein divided by the number of replicates) for a treatment equals the proportion for the control. Failure of the null hypothesis identified 12 significantly up-regulated proteins based on peptide data from treated samples in the absence of sufficient data from controls.

For the remaining 859 proteins (1240 – 303 – 78) where peptide observations satisfied the 2 + 2 rule and included data from controls, peptide abundances were fit with a mixed effects linear statistical model [81] to identify statistically significantly up- and down-regulated proteins at a false discovery rate of 5%, while simultaneously allowing for peptide measurability and LC column effects (see Instrument Effects above). This analysis yielded 383 proteins with significantly different concentrations relative to controls. When combined with the 12 proteins judged to be up-regulated based on peptides observed in treated samples but not controls, we obtain 395 out of 1240 proteins with abundances significantly affected by one or more treatments.

Peptides that are unique to a single protein will have equal molar concentrations in the digested mixture. After instrument effects have been removed (peptide measurability, column effects, etc.) a simple average of their observed abundances provides an estimate of the relative concentration of the protein that they identify. On the other hand, degenerate peptides may come from proteins that are affected differently by a given treatment, which introduces an element of uncertainty if they are included in estimates of protein abundance. For this reason, of the 383 proteins judged to be up- or down-regulated by our mixed-effects statistical model, we are most confident about the 131 differentially abundant proteins that were identified by unique peptides and least confident about the 160 differentially abundant proteins that were identified by degenerate peptides only (These part of data are processed separately as illustrated in Chapter 3 Section 3.3). The remaining 92 differentially abundant proteins came from a set of 197 proteins identified by a mixture of unique and degenerate peptides. An alternative method for assessing the abundance of these proteins, which eliminates the ambiguity associated with degenerate peptides, is discussed in Chapter 5. Based on unique peptides only, 62 of the 197 proteins identified by a mixture of unique and degenerate peptides were judged to have significantly different abundance in treated and controlled samples.

4.1.2.4 Validation of MS results by immunoblotting

In an effort to validate our statistical method to identify up- and down-regulated proteins, immuno-blotting procedures were employed. For this, commercially available antibodies were obtained for 6 proteins. Antibodies were selected based on their

availability and the magnitude of the abundance change. Both up- and down-regulated proteins were selected as well as a protein not observed in the control group. Initially we used the slot blotting technique and visual inspection to rapidly determine whether the immunological approach was in agreement with the MS statistical data. In all 6 cases there was good qualitative agreement between the two methods, i.e., proteins identified as up-regulated by MS also appeared up-regulated on the slot blots and vice versa.

We also wanted to obtain detailed quantitative information on a smaller number of proteins and to compare these results to those from the MS analyses. Western blots were chosen for these analyses because it is generally accepted that they provide more reliable quantitative data than slot blots. Results for Surfactant Protein-D clearly demonstrate that this lung protein is highly up-regulated by all three treatments and that there is good quantitative agreement between the Western blot and the MS statistical method. Western blot analyses for Cathepsin D indicate clear up-regulation in groups receiving smoke treatments which is entirely consistent with the MS statistical results. This points to the importance of statistical analysis for cases where data is missing for the control group. Similar agreement between the MS statistical method and Western blots were obtained for lymphocyte specific protein-1 and haptoglobin. These results provide compelling evidence that our MS statistical method applied to MS data correctly identifies proteins with altered abundance.

4.1.2.5 Summary of results

Proteoviz combined the 62 class-3 proteins judged to be differentially abundant based on unique peptides with the 131 differentially abundant class-1 and -2 proteins

together with the 12 proteins judged to be up-regulated based on peptides observed in treated samples but not controls, and resulted in a total of 205 statistically significant up- or down-regulated proteins. Venn diagram of the distribution of these 205 proteins in the three treatments is shown in Figure 4.5, which indicates the high degree to which proteins with significantly altered concentrations are common to all 3 treatment groups (99 out of 205). A large number of differentially-altered proteins are also common to the CS and combination CS/LPS treatments (39 out of 205). Of the remaining 67 proteins, 12, 15, and 26 are unique to the LPS, CS, and CS/LPS treatments, respectively. Table 4.1 shows the number of up- and down-regulated proteins by treatment groups and class. Proteins identified by a single unique peptide (class 2) contribute fewer statistically-significant up- and down-regulated proteins even though this class contains more proteins than classes 1 or 3 (360 vs 263 and 197). This result points out the advantage of having multiple mass tags for abundance estimates by statistically rigorous methods.

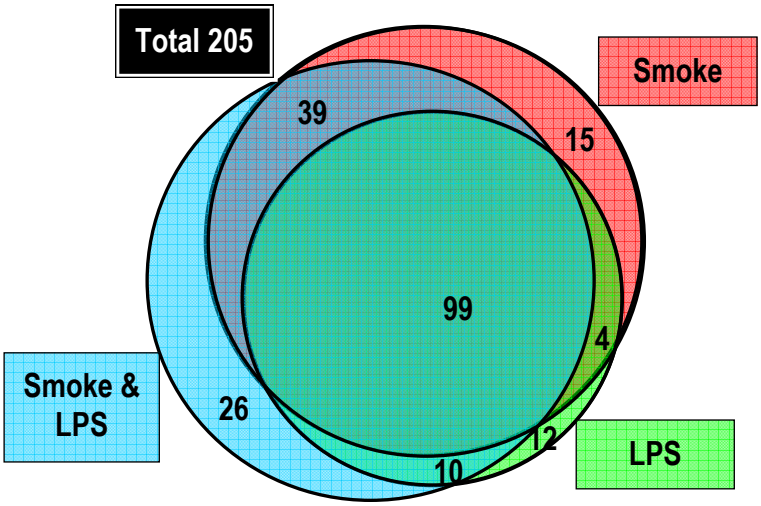

Figure 4.5: Venn diagram of proteins identified by at least one unique peptide and with statistically significantly altered abundance relative to controls at a 5% false discovery rate.

Table 4.1: Shown are the number of proteins with significantly altered abundance (5% false discovery rate) classified by Protein Prophet as being identified by unique or degenerate peptides. The same protein may be altered by one or more treatments.

Treatment	[a]Class 1	[b]Class 2	[c]Class 3	[d]Absent in Controls
Total proteins (205)	101	30	62	12
LPS up regulated	54	10	29	4
LPS down regulated	15	3	10	0
SMK up regulated	66	13	22	10
SMK down regulated	19	8	19	0
LPS+SMK up regulated	66	10	30	12
LPS+SMK down regulated	22	11	23	0

[a]identified by multiple unique peptides

[b]identified by single unique peptide

[c]identified by unique and degenerate peptides; abundance based on unique peptides

[d]observed in treated samples but not controls

Our results demonstrate that label-free FTICR-MS methods can be applied to complex protein mixtures to determine the subset of proteins that are significantly altered in abundance by biological treatment. Advanced statistical analysis to separate treatment effects from instrument effects is an essential component of this approach. Missing peptide data causes unbalance in treatment datasets and necessitates the use of maximum likelihood methods of variance analysis to accurately access which proteins have

statistically significant abundance change relative controls. Identification of up- and down-regulated proteins is the cornerstone of biological interpretation of treatment effects. We feel that statistically rigorous methods like those used in this work are a major improvement over arbitrarily-chosen thresholds of abundance change for identifying these critical features of global protein profiling. Exclusion of degenerate peptides had a significant impact on the number of proteins to which our mixed-effects statistical model could be applied; hence, the development of techniques discussed in Chapter 5 to include degenerate peptides in the assessment of statistically significant protein-abundance change make an important contribution to proteomics.

4.1.2.6 Biological interpretation

Analysis of significantly-altered protein abundance data by MetaCore ranked cell motility and cell adhesion among the important processes induced by the treatments (Figure 4.6).

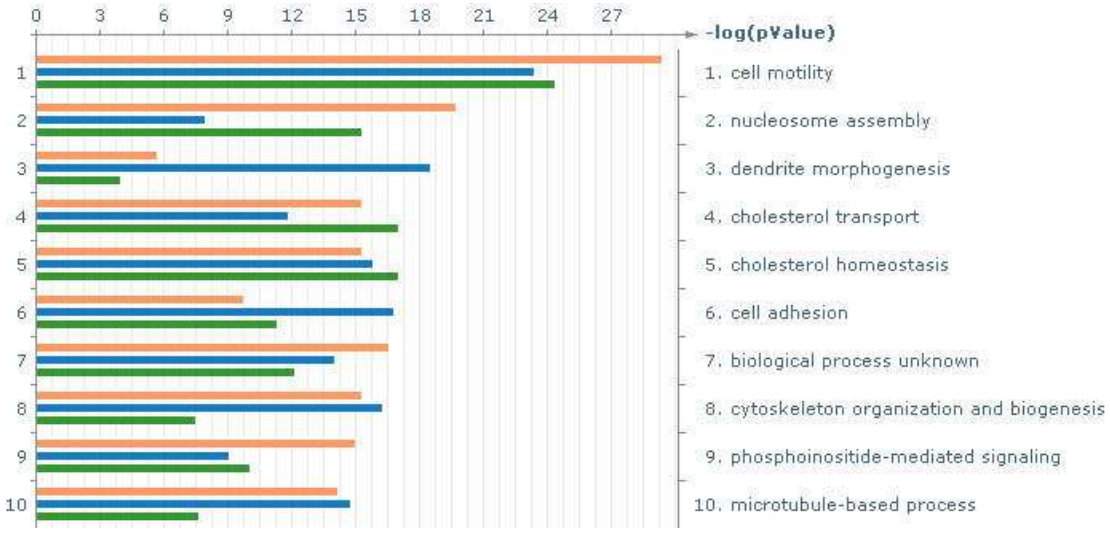

Figure 4.6: MetaCore output showing the distribution of p-values for the top-10 processes associated with up- and down-regulated proteins from treatments in the

mouse-lung toxicology study. From up to down, the three bars in each set represent LPS, SMK, and the combined LPS plus SMK treatment, respectively.

Cell motility is one of the key processes mediating the inflammatory and immune response mediation common to all the three treatment groups. Although some of the differentially-altered proteins involved in cell motility are common to the three treatments, Figure 4.6 suggests some differences exist that make the process most dominant for the combined CS/LPS treatment. After combining our proteomic data with the cytokine data from the same study [84, 85], the MetaCore software was used to build signaling networks based upon highest representation of the cell motility process, which are shown in Figures 4.7, 4.8 and 4.9 for LPS, CS, and combined CS/LPS treatment groups, respectively. Heavy blue line connects proteins directly involved in cell motility based on information in the MetaCoreTM database from published literature. Light lines show known interactions of these proteins with other proteins in the network. Proteins in the network are arranged by subcellular location. Taken together, these networks provide insight into the similarities and differences between the inflammatory responses of the lung to CS and LPS (and combined) exposure.

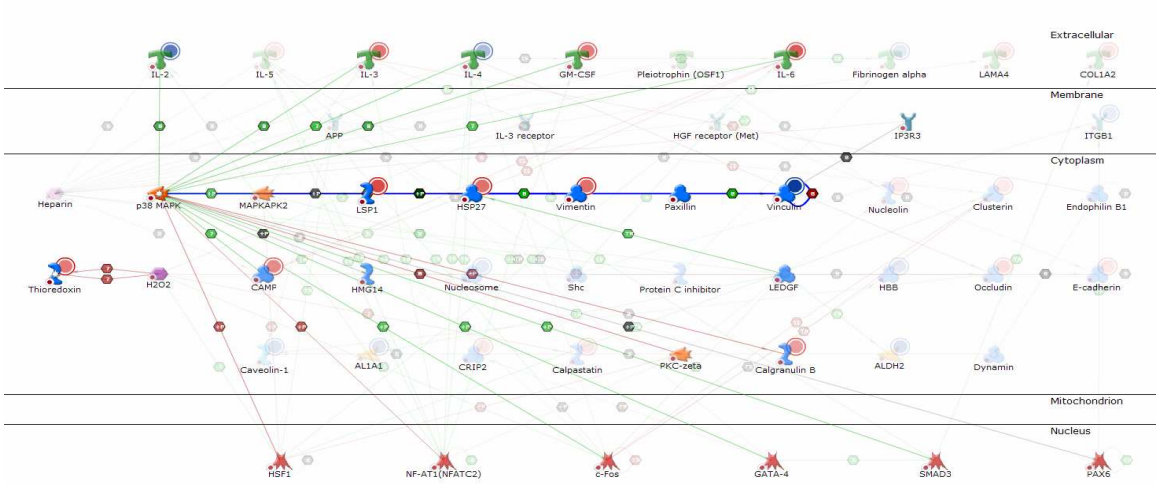

Figure 4.7: Network with the highest representation of the cell motility process built for proteins up-regulated (red circles) or down-regulated (blue circles) by treatment with LPS.

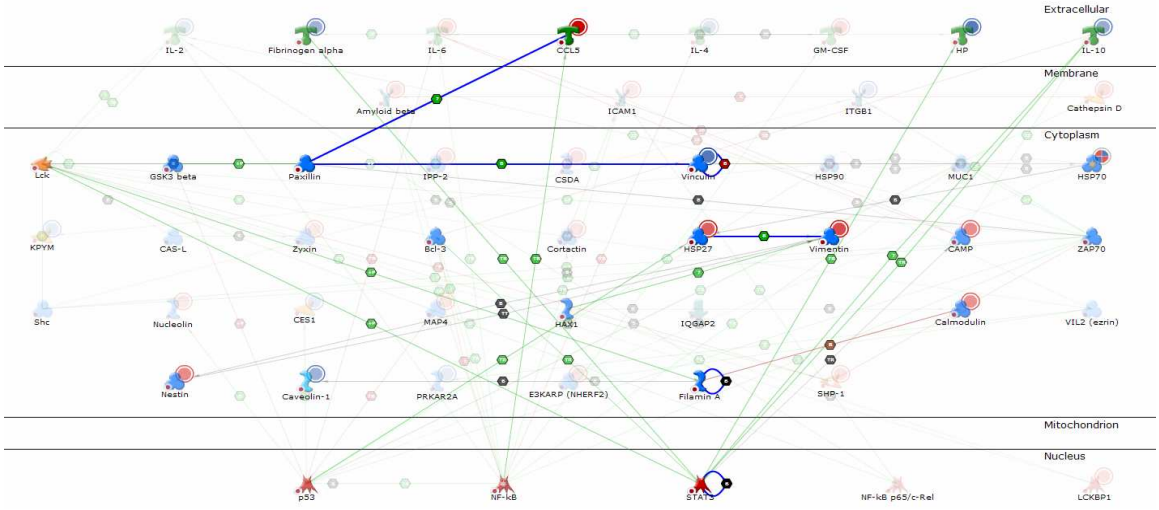

Figure 4.8: Network with the highest representation of the cell motility process built for proteins up-regulated (red circles) or down-regulated (blue circles) by treatment with cigarette smoke (CS).

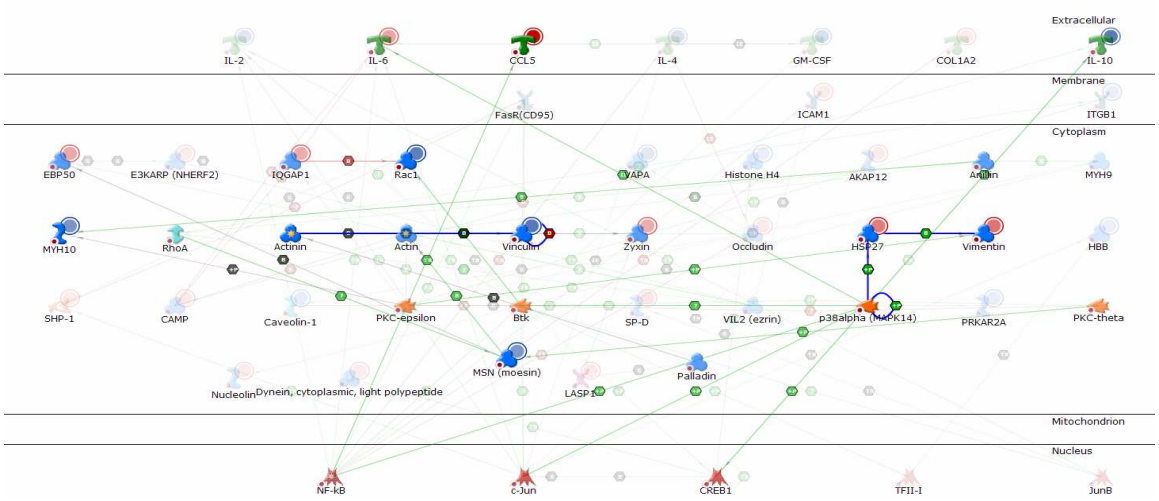

Figure 4.9: Network with the highest representation of the cell motility process built for proteins up-regulated (red circles) or down-regulated (blue circles) by treatment with a combination of LPS and cigarette smoke.

4.1.3 Evaluation of Proteoviz dataflow

The greatest challenge in using new high-throughput technologies for protein profiling is devising ways to extract the full meaning and implications of the data to facilitate data mining. By conducting this COPD study according to our dataflow and processing the data using Proteoviz, we successfully obtained the ultimate dataset in the ideal format to follow up the discovery and characterization of important biological and disease pathways. Reliable, fast and semi-automated, it is proved to be a good helper for proteomics analysis and research work.

4.1.4 Conclusion

In summary we describe the results of a MS-based proteomics study to identify mouse lung proteins with abundance changes attributable to inhalation exposure to CS,

LPS, or combined treatment. The approach includes: 1) FTICR-MS which provides very high mass accuracy for peptide identifications, 2) Protein Prophet to minimize redundancies in the number of proteins identified, 3) A mixed effects statistical model along with a 5% false discovery rate to identify proteins with abundance changes relative to controls, and 4) MetaCore software analysis to reveal biological processes that are common or unique to the treatments. Using this approach we found 205 up- and down-regulated proteins. Approximately one half of these proteins were common to all treatment groups while a smaller number of proteins were treatment-group specific. Using a powerful pathway mapping tool, the most common biological processes identified could be related to inflammation and immune response. These results demonstrate the importance of rigorous statistical evaluation of MS data, provide a prototype for data analysis workflow, and allow for interpretation of the biological processes involved in the inflammatory response to CS/LPS combined treatment, a potential animal model of COPD.

4.2 RIGI study

Radiation-induced genome instability (RIGI) is a response to radiation exposure [86-88] in which the progeny of surviving cells exhibit increased frequency of chromosomal changes many generations after the initial insult. Since genomic instability is believed to be a major factor in tumor promotion [89], understanding the mechanisms that initiate the perpetuate RIGI is of great importance.

Persistently elevated oxidative stress accompanying RIGI and the ability of free-radical scavengers, given before irradiation, to reduce the incidence of instability suggest

that radiation induced alterations to mitochondrial function likely play a role in RIGI. To further elucidate this mechanism, we performed high-throughput quantitative mass spectrometry on samples enriched in mitochondrial proteins from three chromosomally-unstable GM10115 Chinese-hamster-ovary cell lines and their stable parental cell line.

4.2.1 Materials and methods

4.2.1.1 Cell lines

Parental GM10115 hamster human hybrid cells and three cell clones independently derived from single GM10115 cells surviving exposure to ionizing radiation, LS-12, CS-9, and 115 were used for these studies.

4.2.1.2 Sample preparation

Both stable and unstable cell lines were grown in DMEM media to confluence in T150 flasks, harvested by centrifugation and washed. Flasks were pooled to give approximately 6×10^7 cells for each cell line and mitochondria isolated using the Qiagen mitochondrial isolation kit. Isolated mitochondrial proteins were resuspended in bicarbonate buffer and digested with trypsin. Digested peptides were further purified and stored at -80C prior to mass spectrometry analysis. Completeness of the tryptic digestion was confirmed using 4-12% Bis-Tris PAGE gel separation.

4.2.1.3 LC-MS/MS spectrometry

The trypsin digests of purified mitochondrial proteins were each analyzed 3 times by liquid chromatography-tandem mass spectrometry as follows. The digests were

diluted to an approximate concentration in 0.1% formic acid and 2 % acetonitrile (solvent A). The peptides samples were then separated on a capillary column over a gradient from solvent B (0.1% formic acid, 95% acetonitrile) on a liquid chromatography system. The eluting peptides were analyzed by a linear ion trap mass spectrometer equipped with a dynamic nanospray probe. MS and MS/MS spectra were acquired.

4.2.1.4 Data analysis

As illustrated in Chapter 2 the previous work with CHO cells [90, 91], we used both mouse [92] and human [93] protein databases to match in silico LC-MS/MS spectra. This approach to cross-species protein identification could fail due to sequence differences between hamster proteins and their human and mouse homologs; however, due to the small phylogenetic distance between mouse and hamster, we think such failures are minimal. Limitations of the alternative approach of de novo peptide sequencing followed by homology searches, which is far more computationally intense, have been discussed by Habermann et al [94]. The searches were done by using Sorcerer v2.0 (SageN research, San Jose, CA) which combines SEQUEST scoring algorithm and TPP (Trans-Proteomics Pipeline, Institute of System Biology, WA) validating algorithm. SEQUEST output was validated by Protein Prophet to deal with issues of partial coverage and database redundancy. A Protein Prophet probability greater than 0.5 and number of unique peptides in proteins larger than 1 were used to filter the protein identification results. Spectral count, the total number of MS/MS spectra matched to peptides that identify a protein, was used as an indicator of that protein's abundance. Dunnett's test [95] for comparing each of several experimental means with a control

mean was used to infer significantly difference protein abundance between unstable cells lines and the stable parental GM10115 cell line.

4.2.2 Detailed implementation issues

There are several unique characteristics of RIGI study, so although the overall dataflow shares the same idea with COPD study, data processing in RIGI has slight differences. During implementation, we tackled these issues one by one. Figure 4.10 outlines the key data processing steps involved in the dataflow.

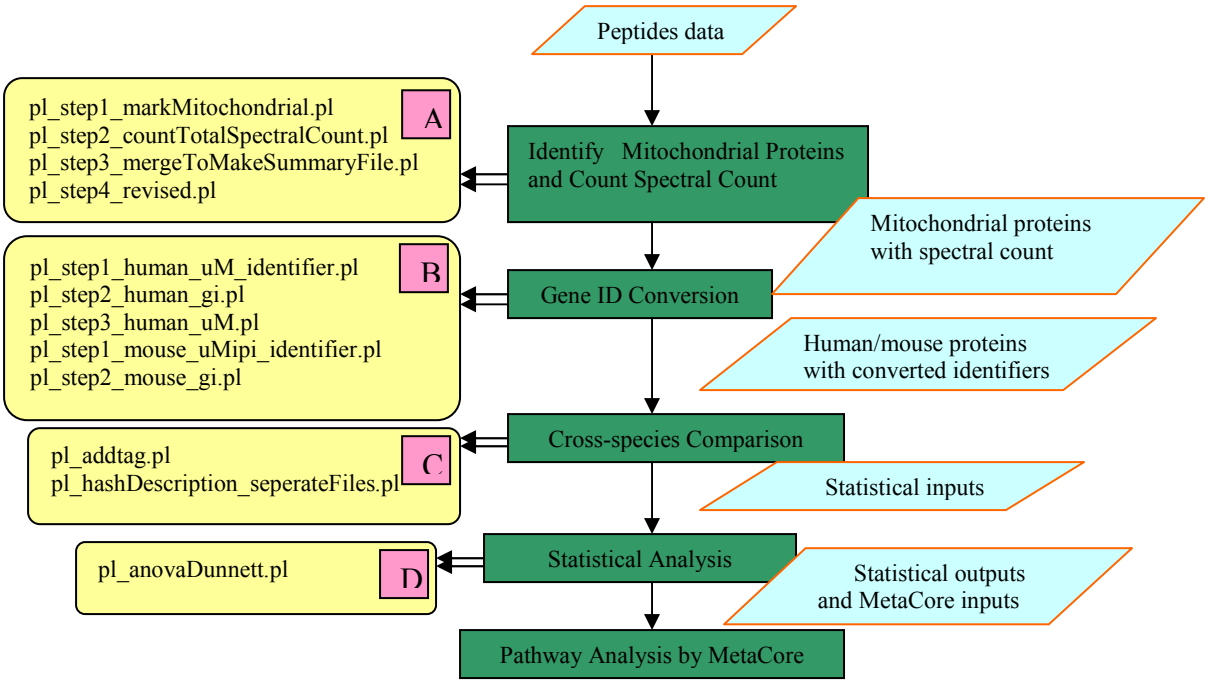

Figure 4.10: Outline of the major data processing steps.

As shown in Figure 4.10 Block A, we identify mitochondrial proteins and calculate its spectral count by searching separate mouse and human databases are the first step in characterizing the mitochondrial proteome of CHO cell lines. Unlike COPD

where we can search against a mouse.fasta protein database for peptide identification, in the RIGI study, we used Chinese-hamster-ovary cell lines, for which a protein database is not available. The small phylogenetic separation between mouse, hamster and human allows the use of protein databases of closely related species to obtain cross-species protein identification; however we expect substantial protein redundancy in the hamster proteins identified by searches against mouse and human databases. Identifying the common proteins between human and mouse database searches is the main challenge of cross-species manipulation. Our scripts helped to covert protein ID to a uniform identifier and by constructing a keyword list, the hash table implementation enables us to find out the common proteins between the two searches based on their fasta annotation (Figure 4.10 Block B and C). This semi-automated procedure for removing redundancy in corss-species protein identification is a substantial improvement over the manual procedure used in the previous studies with CHO described in Chapter 2.

The use of Dunnett's test to infer significantly difference protein abundance between unstable cells lines and the stable controlled GM10115 cell line and the application of MetaCore pathway analysis for biological interpretation of differentially abundant proteins are further improvement over previous work (Figure 4.10 Block D).

4.2.3 Results and discussion

Three samples from each of the four cell lines were subjected to LC-MS/MS spectroscopy and data analysis. Table 4.2 shows the average number of peptides and proteins (total and mitochondrial) identified by searching the mouse and human databases to match spectra obtained with samples from the GM10115 parental cell line. Similar

results were obtained for the 3 unstable cell lines. Mouse database searches consistently yielded a slightly greater number of mitochondrial protein identifications.

Table 4.2: Number of peptide and protein identifications in parental GM10115 cells

	Mouse database	Human database
Number of peptides	898 ± 84	763 ± 65
Number of proteins	333 ± 21	324 ± 19
Number of mitochondrial proteins	93 ± 5	83 ± 3

Many mouse-human homologs can be identified from FASTA annotation without carrying out sequence comparison. We used this approach to eliminate redundancy in the mitochondrial proteins identified by SEQUEST searches of the mouse and human protein databases. Usually, about half of the proteins identified by searching the human database were homologs to proteins found in the mouse database. This overlap reduced the number of distinct mitochondrial proteins identified to about 100 for a typical cell line.

Spectral count is a property of peptides; namely, the number of MS/MS spectra acquired that identify a particular peptide. The sum of the spectral counts for all the peptides that identified a protein was used to estimate the abundance of that protein in an unstable cell line relative the stable parental control. This approach assigns larger total spectral count to proteins identified by multiple peptides and makes their relative abundance estimates less sensitive to random experimental errors. Dunnett's test for protein abundances in unstable cells that are significant difference from control requires a non-zero total spectral count for all 4 cell lines. This requirement eliminated some

proteins identified by only one or a few peptides; however, these proteins were more likely to be false positives than those included in the statistical analysis of differential abundance.

Table 4.3 list those mitochondrial proteins judged to be significantly up- or down-regulated in at least one of the unstable cell lines. Only one of the 15 proteins listed in Table 4.3 was found to be significantly modified in more than one of the 3 unstable cell lines. Acetyl-CoA-acetyltransferase (also known as thiolase), a key component of the tricarboxylic acid (TCA) cycle, is down regulated in both the LS-12 and 115 unstable cell lines. 115 is also character by down-regulation of a protein of unknown function that binds to the Q subcomponent of complement component 1 and by up-regulation of a transmembrane protein that contains an EF-hand domain and, consequently, is most likely associated with calcium-ion (Ca2+) transport.

Among the unstable cell lines investigated in this study, LS-12 is the best characterized at the cellular level. Kim et al. [96] observed that both the state 3 and the uncoupled respiration rates of LS-12 were reduced by about 40% relative to the parental GM10115 cells. They also detected a 30% reduction in the activity of complex IV (cytochrome c oxidase), the last step in the electron transfer chain (ETC). This observation could explain the reduced state-3 respiratory rate and the persistently elevated levels of ROS in LS-12, since blocks to the ETC are known to increase the production of ROS by mitochondria.

The 7 mitochondrial proteins that we observed to have significantly altered abundance in LS-12 relative to the parental GM10115 cell line are listed in Table 4.4 along with their catalytic activity and the metabolic pathway in which they function. In

Table 4.3: Proteins abundance in unstable cells relative to stable parental control

Protein	LS-12/GM	CS-9/GM	115/GM
acetyl-CoA-acetyltransferase	0.55*	0.84	0.77*
aconitate hydratase (1)	0.53*	0.84	0.96
ATP synthase (beta)	1	1.15*	1
calcium-binding (ARALAR2)	1.2	1.86*	1.44
citrate synthase	0.65*	0.95	0.89
complement component	0.6	0.8	0.53*
cytochrome c oxidase (subunit 5A)	1	1.63*	1.25
enoyl-CoA hydratase (alpha)	0.52*	0.96	1
isocitrate dehydrogenase (NADP+)	0.58*	0.92	1
isocitrate dehydrogenase (alpha)	0.99	1.18*	1.12
isocitrate dehydrogenase (isoform 2)	0.39*	0.94	1.17
Leucine zipper-EF-hand transmembrane	1.42	1	1.67*
pyruvate dehydrogenase (beta)	0.31*	0.89	0.74
succinate dehydrogenase (flavoprotein)	0.82	1.30*	0.81
glutamate dehydrogenase	0.56*	0.96	0.95

* Statistically significant at 95% confidence

all 7 cases, the proteins are down-regulated and involved in fatty acid or carbohydrate metabolism. As mentioned by Kim et al. altered abundance and/or activity of TCA enzymes can be the cause of reduced state-3 respiration; however, their observation of reduced cytochrome c oxidase activity leads them to suspect the ETC. The coupling

between central carbon metabolism and oxidative phosphorylation makes it difficult to discern cause from effect; nevertheless, mutations in mitochondrial DNA that cause blockage in the ETC and metabolic shifts away from the TCA-cycle have been reported for budding yeast and C. elegans [97].

Table 4.4: Down-regulated mitochondrial proteins in LS-12 cells

Protein	catalytic activity	pathway
acetyl-CoA-acetyltransferase	Acyl-CoA + acetyl-CoA = CoA + 3-oxoacyl-CoA	lipid and fatty acid metaboloism
aconitate hydratase	Citrate = cis-aconitate + H2O	tricarboxylic acid pathway
citrate synthase	Acetyl-CoA + H2O + oxaloacetate = citrate + CoA	tricarboxylic acid pathway
enoyl-CoA hydratase (alpha)	(3S)-3-hydroxyacyl-CoA = trans-2-enoyl-CoA + H2O	fatty acid beta-oxidation cycle
isocitrate dehydrogenase (isoform 2)	isocitrate + NADP(+) = 2-oxoglutarate + CO2 + NADPH	interacts with pyruvate dehydrogenase
pyruvate dehydrogenase (beta)	conversion of pyruvate to acetyl-CoA and CO2	intermediary metabolism
glutamate dehydrogenase	L-glutamate + H2O + NADP(+) = 2-oxoglutarate + NH3 + NADPH	catabolism of glutamate

Table 4.5 list the 5 mitochondrial proteins we observed to have significantly altered abundance in CS-9 unstable cell line relative to parental GM10115 cells. This protein profile of dysfunctional mitochondria contrasts sharply with that described above for LS-12. In this case we observed up-regulated proteins, none of which have significantly altered abundance in LS-12 unstable cells. Three of the up-regulated mitochondrial proteins in CS-9, ATP synthase, succinate dehydrogenase and cytochrome c oxidase, are directly associated oxidative phosphorylation and ATP production. Like isocitrate dehydrongenase, succinate dehydrogenase is also part of the TCA-cycle.

Table 4.5: Up-regulated mitochondrial proteins in CS-9 cells

Protein	catalytic activity	pathway
ATP synthase(beta)	ATP + H2O + H(+)(In) = ADP + phosphate + H(+)(Out)	ATP production in presence of H+ gradient
calcium-binding (ARALAR2)	not an enzyme	aspartate amd glutamate carrier
cyctochrome c oxidase (subunit 5A)	4 ferrocytochrome c + O2 = 4 ferricytochrome c + 2 H2O	terminal oxidase in electron transport
isocitrate dehydrogenase (alpha)	isocitrate + NADP(+) = 2-oxoglutarate + CO2 + NADPH	intermediary metabolism and energy production
succinate dehydrogenase	succinate + ubiquinone = fumarate + ubiquinol	component of complex II

Characterization of CS-9 at the cellular level is not as complete as for LS-12. Unpublished work by Spitz and coworkers indicates that, like LS-12, CS-9 has persistently elevated ROS and compromised antioxidant capacity. Our results suggests that different mechanisms may be responsible for the elevated levels of ROS in LS-12 and CS-9. One possible interpretation of our data on CS-9 is that the respiration rate is higher than in the parental GM01115 cells, so that a similar leakage rate could produce more ROS.

4.2.4 Conclusions

Research on dysfunctional mitochondria has increased in recent years due to the discovery of their role in human diseases that include cancer, diabetes, neurodegeneration and cardiomyopathy. Recognition of their role in both the direct and non-targeted responses of cells to radiation exposure has also increased. In both human disease and radiation response, the role of mitochondria can be explained in general terms by the model "Mitochondrial Threshold Effect Theory" [98]. This model is based on the hypothesis that cells cope with a certain degree of mitochondrial dysfunction by compensatory mechanisms that support viability. Increased mitochondrial mass after radiation exposure is an example of this principle that may have the undesirable long-term effects of persistent oxidative stress and genome instability; however, this mechanism does not apply to the LS-12 cell line, since it does not show an increase in the number of mitochondria. The distinct mitochondrial-protein profiles of genome unstable cell lines observed in this study suggest that a variety of coping mechanisms are available

to cells with compromised mitochondrial function due to radiation exposure. Understanding these compensatory mechanisms and their thresholds for failure are an important new area of research in radiation biology.

Chapter 5

Degenerate Peptides

5.1 Degeneracy related protein abundance estimate issues

Assessment of differential protein expression from the observed properties of detected peptides is the primary goal of label-free shotgun proteomics. The abundance observed for unique peptide originates from its identified protein only. However, the abundance observed for degenerate peptides may be due to contributions from multiple proteins in a biological sample that are affected differently by a given treatment. Consequently, including degenerate peptides in estimates of protein abundance may lead to erroneous results for the effect of a treatment on protein expression levels.

Excluding degenerate peptides eliminates this ambiguity but may significantly decrease the number of proteins for which abundance estimates can be made, especially when degenerate peptides are in a large fraction. So when degenerate peptides show up in the identification of a protein, it is helpful to develop a strategy to evaluate the contribution they give to an estimate of the protein's abundance and make a judgment on whether or not to include them in the protein's abundance estimate. At the early stage, we start our exploration by an approach based on the following assumption: *If a protein is*

identified by multiple peptides, the abundances of the peptides in samples from a treatment group relative to its abundance in samples from control should all be consistent with each other. We know that unique peptides have this property because they will be present at the same concentration as the protein from which they were derived by proteolysis; hence we require it to be true for degenerate peptides also.

To implement this approach, we can pick up all the unique peptides that identify a protein, and calculate their ratio of abundance as a reference. Then we calculate the ratio of abundance of the degenerate peptides that identify this protein. Compare each of them with the ratio of unique peptides. If the ratios are consistent with each other to a specified level of statistical confidence, we can keep the degenerate peptide for protein's abundance estimate; otherwise, we discard it.

By this way, we can quantitatively evaluate the contribution of degenerate peptides to a protein's abundance estimate. Comparing with previous strategy, which eliminates all the degenerate peptides data, we can achieve a larger dataset by introducing valuable degenerate peptides data and discarding only the degenerate peptides whose abundances are influenced by multiple proteins that respond differently to a given treatment.

The above approach can be applied without considering any biological relationship among the proteins that contain a degenerate peptide. It is purely empirical and only requires sufficient data on unique peptides to define an expected pattern of abundance change with treatments relative to controls. Nevertheless, when we investigated the biological function and annotation of proteins that are identified by a degenerate peptide, we often found that they belong to a family of proteins, all members

of which have similar amino-acid sequence and biological function. Sequence similarity increases the likelihood peptide degeneracy (i.e. a peptide found in one family member is likely to be found in all family members). Functional similarity raises the possibility that treatments may affect all of the family members in a same way. When this is the case, it is reasonable to use both unique and degenerate peptides to discover the correlated effect of treatments on the biologically-related proteins. So we want to try a higher level analysis of the degenerate peptides that endeavors to discover the effect of treatments on the whole family of proteins rather than its individual members. This reasoning leads to the following assumption concerning the use of degenerate peptides in protein abundance estimation: *If all of the proteins that a peptide identifies are in the same family, we treat it as a unique peptide for estimation of the effects of treatments on the family of proteins*.

Based on this assumption, we developed a more biologically based method to include degenerate peptides in protein abundance estimates. The description of proteins in the database used by SEQUEST to interpret LC-MS/MS spectra is usually sufficient to define protein families. In most cases a family consists of isoforms of the same protein. After a protein family has been composed, we can investigate the degenerate peptides one by one to determine if the source of degeneracy is just different family members and form the final peptide abundance dataset for the subsequent protein's abundance estimate and statistical significance analysis.

5.2 The computational method

The above hypotheses form the basis for us to develop our closure-family method to solve degenerate peptides problem in protein abundance estimates. In this section we

describe this method of including degenerate peptides in differential expression analysis for protein profiling using label-free proteomics data. It starts with finding groups of proteins that have a biological relationship and embody all of the degeneracy of peptides that identify group members, then tests for correlated abundance changes in the group due to treatments. Four steps are involved in this computational method.

5.2.1 Finding peptide-degeneracy closure groups

The first step in our method is to identify groups of proteins with closure on degenerate peptides (i.e. the group contains all the proteins responsible for the degeneracy of any peptide that identifies a group member). If we define the level of degeneracy as the number of proteins that contain an observed peptide, then the condition for closure is that the number of occurrences of every peptide in a group must equal its level of degeneracy.

5.2.2 Finding closure groups with biologically related proteins

For many peptide-degeneracy closure groups, the FASTA annotation of proteins in the group is sufficient to explain their biological relationship. An example of groups of this type is shown in Table 5.1, where the relationship is a predicted similarity to different isoforms of laminin alpha 5.

In some cases, like that shown in Table 5.2, key words in the FASTA annotation suggest a biological relationship but are not sufficient to determine the putative relationship. Ubiquilin 2 is probably biologically related to ubiqulin 1 but more complete annotations are required to confirm this hypothesis. In cases like this, we assumed that

the FASTA annotation was evidence for a biological relationship among proteins of a peptide-degeneracy closure group.

In a relatively small number of cases, FASTA annotation suggested that the proteins in a peptide-degeneracy closure group have different biological functions. An example of this type is shown in Table 5.3, where the degenerate peptide AAIDWFDGK is found in both pigpen and TAF15. Cases of this type are flagged for further analysis to assess the likelihood that more than one protein was contributing to the observed abundance of the degenerate peptide.

Table 5.1: Proteins in the Laminin peptide-degeneracy closure group

Protein description: similar to laminin alpha 5 isoform	Peptide	Level of Degeneracy
1	AVEASNAYSSILQAVQAAEDAAGQALR	1
1	GQLQLVEGNFR	1
1	AHPVSNAIDGTER	2
1	ATGDPWLTDGSYLDGSGFAR	2
3	ATGDPWLTDGSYLDGSGFAR	2
9	AHPVSNAIDGTER	2

Table 5.2: Proteins in the Ubiquilin peptide-degeneracy closure group

Protein Description	Peptide	Level of Degeneracy
ubiquilin 1 isoform 1	QQLPTFLQQMQNPDTLSAMSNPR	1
ubiquilin 1 isoform 1	EKEEFAVPENSSVQQFK	1
ubiquilin 1 isoform 2	NQDLALSNLESIPGGYNALR	1
ubiquilin 1 isoform 2	NPEISHMLNNPDIMR	1
ubiquilin 1 isoform 2	FQQQLEQLSAMGFLNR	2
ubiquilin 1 isoform 2	QLIMANPQMQQLIQR	2
ubiquilin 1 isoform 2	ALSNLESIPGGYNALR	2
ubiquilin 2	GPAAAPGAASPPAEPK	1
ubiquilin 2	FQQQLEQLNAMGFLNR	1
ubiquilin 2	ALSNLESIPGGYNALR	2
ubiquilin 2	QLIMANPQMQQLIQR	2
ubiquilin 2	FQQQLEQLSAMGFLNR	2

Table 5.3: Proteins in the Pigpen peptide-degeneracy closure group

Protein Description	Peptide	Level of Degeneracy
Pigpen	TGQPMINLYTDR	1
Pigpen	EFSGNPIK	1
Pigpen	AAIDWFDGK	2
TAF15 RNA polymerase II, TATA box binding protein associated factor	AAIDWFDGK	2

5.2.3 Assessment of common abundance change for proteins in a peptide-degeneracy closure group

Let A_x^t and A_x^c be the observed abundances of peptide x in treatment and control groups, respectively. Consider a peptide-degeneracy closure group consisting of proteins identified by unique peptides *p* and *q*, as well as a degenerate peptide *pq*. The abundances of these peptides in samples from treatment group *t* relative to their abundances in control samples are given by

$$\frac{A_p^t}{A_p^c} = \frac{C_p^t}{C_p^c} = r_p \tag{1}$$

$$\frac{A_q^t}{A_q^c} = \frac{C_q^t}{C_q^c} = r_q \tag{2}$$

$$\frac{A_{pq}^t}{A_{pq}^c} = \frac{C_p^t + C_q^t}{C_p^c + C_q^t} = r_{pq} \tag{3}$$

where A_p^t and A_q^t are the concentrations of proteins p and q in samples from treatment t and similarly C_p^t and C_q^t are their concentrations in control samples. Equations (1-3) are based on a model of observed peptide abundances in which the variation of detection sensitivity for peptides of different amino acid sequence is contained in multiplicity factors that are the same for samples from treatment and control groups; consequently, these factors cancel in peptide abundances relative to controls. In this model, one can easily show that if $r_p = r_q = r$, then $r_{pq} = r$. This result generalizes for any combination of proteins responsible for a peptide's degeneracy; hence, we can include both unique and

degenerate peptides in a one-parameter model for the abundance relative to control of proteins in a peptide-degeneracy closure group.

5.2.4 Statistical analysis

For each treatment *t*, mean abundances relative to control and their standard error are calculated for all peptides (unique and degenerate) identifying proteins in a group with closure on peptide degeneracy. Well-known propagation-of-error formulas [99] are used to transform these quantities into log-space where statistical tests based on the assumption of normally-distributed errors are expected to be more valid. Let ρ_{it} and σ_{it} denote the mean and standard error of the logarithm of peptide abundance relative to control. If detection of proteins in a group defined by closure on peptide degeneracy involved *k* peptides, then the best estimator [99] for the common abundance of proteins in the group is

$$\hat{C}_t = \sum_{i=1}^{k} c_{it}\rho_{it} \tag{4}$$

where

$$c_{it} = \frac{1}{\sigma_{it}^2} \Bigg/ \sum_{i=1}^{k} \frac{1}{\sigma_{it}^2} \tag{5}$$

Well-known results for a linear combinations of means [100] were used to derive a confidence interval for $\hat{C}_t$. Superimposing $\hat{C}_t$ with its confidence interval on a log-space plot of mean peptide abundances relative to control gives a visual indication of how well the model of a common abundance relative to control applies to a group of

biologically related proteins with closure on peptide degeneracy. In addition, an F-test [100] is performed on the null hypothesis of equal means for peptide abundances in a peptide-degeneracy closure group. Failure to reject the null hypothesis at a specified confidence level is another indication of the validity of the model that proteins in a peptide-degeneracy closure group have a common abundance relative to control.

5.3 Results on degeneracy approach

In this section, we use data from the COPD study (mouse-lung toxicology study) to illustrate our approach for including degenerate peptides in estimates of differential protein abundance. Detailed illustration of this study was given in Chapter 4. We briefly introduce this study here again for the purpose of presenting the degeneracy approach.

In that study, lung tissue samples are taken from 5 animals in each of 3 treatment groups and 5 control animals. Two LC/MS injections are prepared from each biological sample. By combining data from different animals and injections, 10 replicates are potentially available to assess a peptide's abundance in a given treatment group or control; however, for some peptides the number of replicates actually seen is far less than 10. In the case illustrated by Figure 5.1, missing data for unique peptides severely limits the application of t-tests to assess the equality of means between treatment and control groups; however, these t-tests that require a minimum of 3 replicates [101] can be carried out on abundance measures for the degenerate peptides.

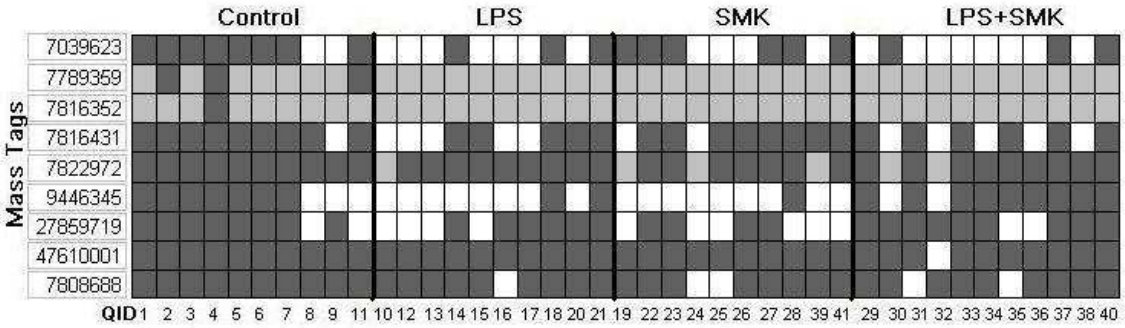

Figure 5.1: The pattern of peptide abundance observations in the ubiquilin peptide-degeneracy closure group (see Table 5.2). Mass Tags are elements of the PMT database used to interpret LC-FTICR mass spectra. White and light-gray squares denote observation of unique and degenerate peptides, respectively. A dark-gray square indicates that the peptide was not observed in the injection.

The ability to use both unique and degenerate peptides in estimates of protein abundance diminishes the impact of missing peptide data and increases the number of proteins for which abundance relative to control can be quantified. Analysis of the proteomic data from the mouse lung-tissue study [102, 103] illustrates this point very well. Processing the LC/MS data by SEQUEST and Protein Prophet reveals 1240 proteins in the NCBI mouse database identified by 3219 peptides with peak areas determined by FTICR-MS [104]. The proportion of proteins identified entirely by degenerate peptides (420/1240) is approximately equal to the proportion of degenerate peptides (1045/3219).

Since the Protein-Prophet confidence scores for the 420 proteins identified by degenerate peptides are generally low, excluding them by using only unique peptides in protein abundance estimates may not be a great loss; however, this is not the full impact

of degenerate peptides. Of the remaining 820 proteins identified by at least one unique peptide, 197 are identified by a mixture of unique and degenerate peptides (referred to as the mixed class) and generally received high confidence scores from Protein Prophet. For this set of proteins, the effects of degenerate peptides are assessed by calculating mean abundances relative to controls for the 3 treatments in the mouse toxicology study with and without degenerate peptides. Excluding degenerate peptides significantly affects abundance estimates for about half of the proteins in the mixed group, with the most obvious effect being the number of proteins with sufficient peptide data to estimate abundance relative to control. By including degenerate peptides, relative abundance can be estimated for 178 of the 197 mixed-class proteins, which decreased to 140 when degenerate peptides are excluded. These findings encourage us to develop a method for including degenerate peptides in differential protein-abundance assessments.

Ambiguity of source is the basic reason for excluding degenerate peptides in protein abundance estimates; hence, it is reasonable to seek a higher-order classification of proteins that eliminates this ambiguity. We call this entity a "peptide-degeneracy closure group" because it includes all proteins associated with a set of degenerate peptides. Initially, we are surprised to see how often proteins in a closure group appeared to have similar function. We found 165 peptide-degeneracy closures groups among the 617 proteins identified in the mouse-lung toxicology study partially or completely by degenerate peptides. Based on FASTA annotation, 143 of these closure groups appears to be composed of biologically related proteins. These 143 peptide-degeneracy closure groups contain 82% of the proteins for which identification involved degenerate peptides. Even though definitive conclusions about biological relationships cannot be based on

FASTA annotation alone, this finding suggests that most of the peptide degeneracy in the data from the mouse-lung toxicology study is among protein family members.

The Transferrin peptide-degeneracy closure group (Table 5.4) is among the 22 closure groups where FASTA annotation suggests that proteins in the group are not biologically related.

Table 5.4: Proteins in the Transferrin peptide-degeneracy closure group

Protein description	Peptide	Level of Degeneracy
transferrin	GDVAFVK	2
melanoma associated antigen p97	GDVAFVK	2
transferrin	SKDFQLFSSPLGK	1
transferrin	LYLGHNYVTAIR	1
transferrin	HTTIFEVLPEK	1
transferrin	YLGAEYMQSVGNMR	1
transferrin	KGTDFQLNQLEGK	1
transferrin	TAGWNIPMGMLYNR	1
transferrin	LGHNYVTAIR	1
transferrin	HQTVLDNTEGK	1
transferrin	DSAFGLLR	1
transferrin	DFQLFSSPLGK	1
transferrin	EEYNGYTGAFR	1
transferrin	LPEGTTPEK	1

Since Transferrin is identified by 12 unique peptides and melanoma associated antigen p97 is identified by only one degenerate peptide, it is not surprising that their Protein-Prophet confidence scores are 1 and 0, respectively. The abundance data for the

degenerate peptide GDVAFVK in samples treated with LPS is consistent with data on the unique peptides that identified Transferrin. Similar results are obtained for the other 2 treatments in the mouse toxicology study and the result shown in Figure 5.2 for the best estimate of Transferrin abundance relative to control is validated by immunoblots.

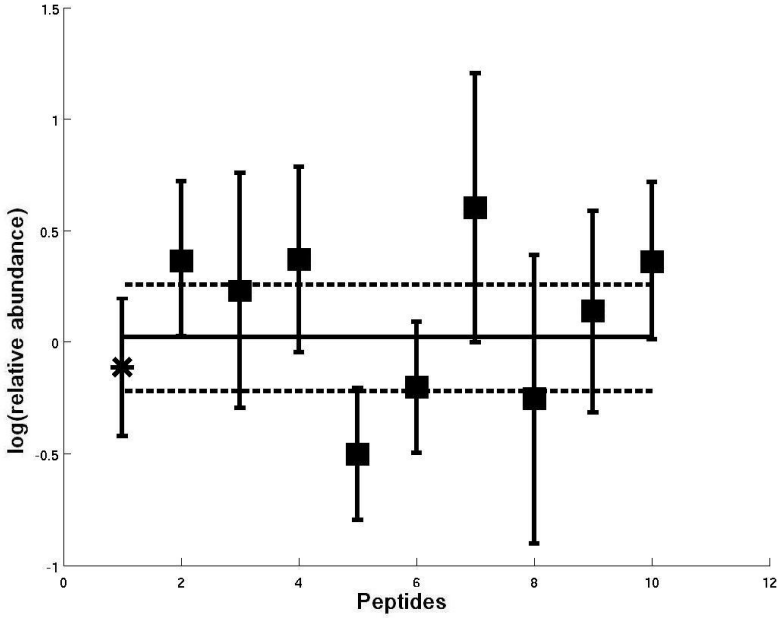

Figure 5.2: Logarithm of the mean relative abundance of peptides identifying proteins in the Transferrin group with closure on peptide degeneracy. Peptides 2 – 11 (squared) uniquely identify Transferrin. Peptide 1 (asterisked) is also found in melanoma associated antigen p97. Dashed lines show the 95% confidence interval on the best estimate of a common relative abundance (solid line) for all 12 peptides.

The F-test for equality of means [100] fails when apply to the data in Figure 5.2; however, the failure is more likely due to scatter in the mean abundances of peptides that uniquely identify Transferrin than to the degenerate peptide (#1 in Figure 5.2), which has

a mean abundance in good agreement with the estimate of a common mean abundance (solid line). To determine how often to expect this type of false negative, we apply the F-test for equality of means to 476 cases in the mouse toxicology study where proteins were identified by multiple unique peptides. Failure of the test in 27% of these cases where peptides are unique to a particular protein suggests that sources of variability other than different protein concentrations could explain failure of the F-test for equality means when it is applied to the abundance of degenerate peptides. Nevertheless, success of the equality of means test adds weight to the conclusion of a common abundance relative to control for biologically-related proteins in a peptide-degeneracy closure group.

Our results for the Transferrin peptide-degeneracy closure group clearly point to the conclusion that peptide GDVAFVK uniquely identifies Transferrin because melanoma associated antigen p97 is probably not in the samples being investigated to any appreciable extent. This example shows that application of Protein-Prophet confidence scores can reduce peptide degeneracy by eliminating false positive protein identifications. We investigate the magnitude of this effect by removing all 381 proteins identified in the mouse-lung toxicology study with a Protein-Prophet confidence score of zero. Since most of the peptides identifying these proteins also identify proteins with nonzero Protein-Prophet confidence scores, only 4 peptide-degeneracy closure groups incur a loss of peptides due to elimination of proteins with zero confidence score. By retaining 23 of the 381 proteins with a Protein-Prophet confidence score of zero, no peptides are lost from any peptide-degeneracy closure group, which means that the data analysis to estimate a common abundance of group members is not affected. In 118 of the 165 peptide-degeneracy closure groups, elimination of proteins with a zero Protein-Prophet

confidence score reduces the level of peptide degeneracy to the point that all the remaining proteins are identified by unique peptides.

Before removal of proteins with Protein-Prophet confidence score of zero, the peptide-degeneracy closure group with FASTA annotations relate to Myosin contains 7 NCBI database entries identified by 11 peptides, two of which are unique to the protein described as “Myosin light chain, regulatory B-like” identified with a Protein-Prophet confidence score of 1. After removal of proteins with Protein-Prophet confidence score of zero, the only other protein in the Myosin closure group has FASTA annotation “Predicted similar to Mysoin regulatory light chain 2” and a Protein-Prophet confidence score of 0.98. Eight of the 11 peptides identifying members of the Myosin peptide-degeneracy closure group are observed in a sufficient number of replicates to carryout our statistical analysis. The equality of means test gives a positive result in this case and Figure 5.3 shows how well the peptide data from mice treated with LPS can be explained by a common relative abundance. The solid horizontal line marks the best one-parameter fit to these mean peptide abundances. The dashed lines bound a 95% confidence interval on the optimum value of this parameter. Since this interval does not include zero, we interpret these data as a statistically significant up regulation of the proteins in this Myosin closure group due to the LPS exposure.

Due to insufficient data, the statistical analysis illustrated by Figure 5.3 can not be carried out for all of the 143 peptide-degeneracy closure groups with biologically related proteins. For the LPS and SMK treatments, 94 closure groups have the minimum requirement of 2 peptide abundance observations in both treatment and control samples to apply the statistical analysis. For the combined LPS+SMK treatment, the statistical

analysis can be performed on 97 peptide-degeneracy closure groups. In total, these groups contain about 60% of the 617 proteins for which the identification involves degenerate peptides. In a majority of the cases where sufficient data are available for statistical analysis, the F-test for equality of means suggests that peptide relative abundances are not significantly different.

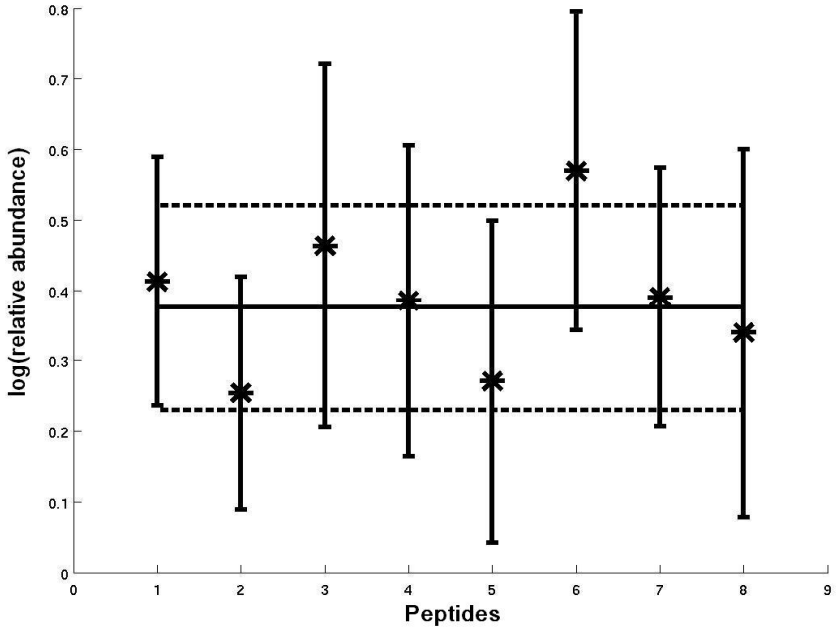

Figure 5.3: One-parameter fit to the logarithm of the mean relative abundances of peptides identifying proteins in the Myosin group with closure on peptide degeneracy. Dashed lines show the 95% confidence interval on the best estimate of a common relative abundance (solid line) for all 8 degenerate peptides (asterisk).

5.4 Significance of the degeneracy approach

Degenerate peptides, a frequent phenomenon in shotgun proteomics, complicate both the identification of proteins and estimates of their relative abundance in biological samples. It is reasonable to have more confidence in proteins identified by unique peptides than degenerate peptides and this is reflected in confidence scores reported by Protein Prophet. The impact of degenerate peptides on the interpretation of shotgun proteomics can usually be reduced by treating proteins with low confidence scores as false positive identifications; however, many researchers are reluctant to take this approach and prefer to retain all identified proteins as a basis for biological interpretation of proteomic data. The concept of peptide-degeneracy closure groups we have presented is a way to deal with degenerate peptides that is somewhat insensitive to the confidence of protein identifications. As the threshold score for confident protein identification increases, the protein composition of peptide-degeneracy closure groups changes but peptide composition is unchanged if degeneracy is due to proteins with both high and low confidence scores.

Our conclusions about peptide-degeneracy closure groups are based on analysis of data from the study of mouse lung tissues where we find that (1) closure groups are most often composed of proteins with related biological function as judged by their FASTA annotation, (2) a one-parameter model of abundance relative to control for all proteins in a closure group is frequently adequate to explain the observed relative peptide abundances, and (3) a large number of proteins identified with low Protein Prophet confidence scores can be rejected as false positives without affecting the peptide abundance data used to discern the affect of treatments on closure groups. Our

conclusions regarding the utility of peptide-degeneracy closure groups will undoubtedly be refined as we apply the concept to additional datasets; however, we are confident that they enable an approach that is an improvement over simply excluding degenerate peptides, which we have found to significantly reduce the number of identified proteins for which differential-abundance analysis can be carried out with statistical inference.

Chapter 6

Future Work

The proceeding chapters presented a seamless dataflow to facilitate the high-throughput proteomics analysis. It deals with data manipulation, degenerate peptides, and pathway analysis in a semi-automated fashion. Some places can be further improved to help us extract the underlying meaning of proteomics data and unravel the biological or disease processes associated with it. These improvements are the targets of our ongoing project.

6.1 Degenerate peptides

Degenerate peptides had a significant impact on the number of proteins to which the assessment of statistically significant protein-abundance change can be performed. We have proposed a peptide-degeneracy closure group approach to include degenerate peptides in protein abundance estimation, and derived a one-parameter statistical model to accomplish this objective. More work could be done in this area. For example, the one-parameter statistical model deals with the degenerate peptide problem in isolation. A

more rigorous statistical model integrating the degenerate peptide issue with peptide measurability and other instrument effects is needed.

6.2 Missing data

The degenerate peptide issue is part of a more general missing-data problem in LC-MS proteomics. Figure 6.1 illustrates a case where many peptides identify a protein, which should allow accurate estimation of protein abundance in samples from control and treatment groups; however, missing data, indicated by the dark squares in the figure, eliminates many of the detected peptides from statistical analysis based on a mixed effects linear model.

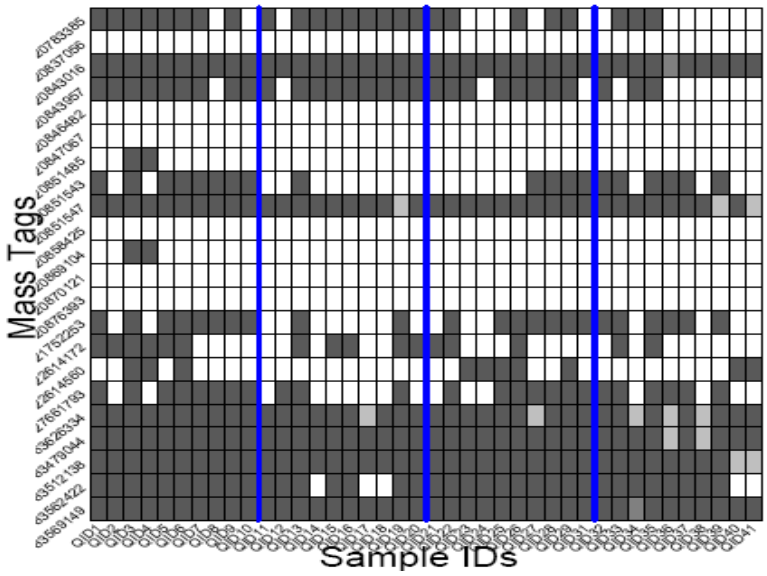

Figure 6.1: Pattern of peptide abundance observations for a protein identified by multiple unique peptides. White indicates samples in which the peptide was observed. Black indicates missing data.

We know that the greatest source of uncertainty in protein abundance estimation comes from the inherent difference in measurability of different peptides that identify a protein. Some of the missing data are true observations due to the fact that the abundance of the peptide in a given protein sample is near the detection limit. But we suspect there are still a great number of missing data that are false observations because the peptide ion-current peak is present in the mass spectrum but misidentified. The strongest evidence for false observations comes from replicate injections into the spectrometer from the same biological sample. Peptides seen with high abundance in one but not all injections of the same sample may be due to choice of parameters used in processing LC-MS/MS data to obtain the PMT database and parameter that control matching of peaks in the FTICR spectrum to entries in the PMT database.

To explore this problem, we plan to look into the COPD dataset, and pair up all the replicate or triplicate injections of a biological sample to determine for each detected peptide whether it is seen in all replicate injections. Figure 6.2 illustrates several distinct conditions that will illuminate the missing-data problem. 1) From abundances of peptides seen in all injections, we can estimate a threshold of abundance to distinguish low abundances from other possible sources of missing data, like incorrect identification. 2) Given this threshold, we can eliminate cases where a peptide seen in one but not all injections is likely due to limited sensitivity. 3) Count the cases where high abundance peptides are missing in one or more replicate injections of the same biological sample.

With the count defined above as a quantitative measure of missing data, we investigate the parameters involved in MS data reduction to determine if there is a correlation between these parameters and the missing data problem. Therefore, we can

formulate criteria to choose suitable parameters to enable minimizing missing data while keeping false-positive peptide identification at an acceptable level.

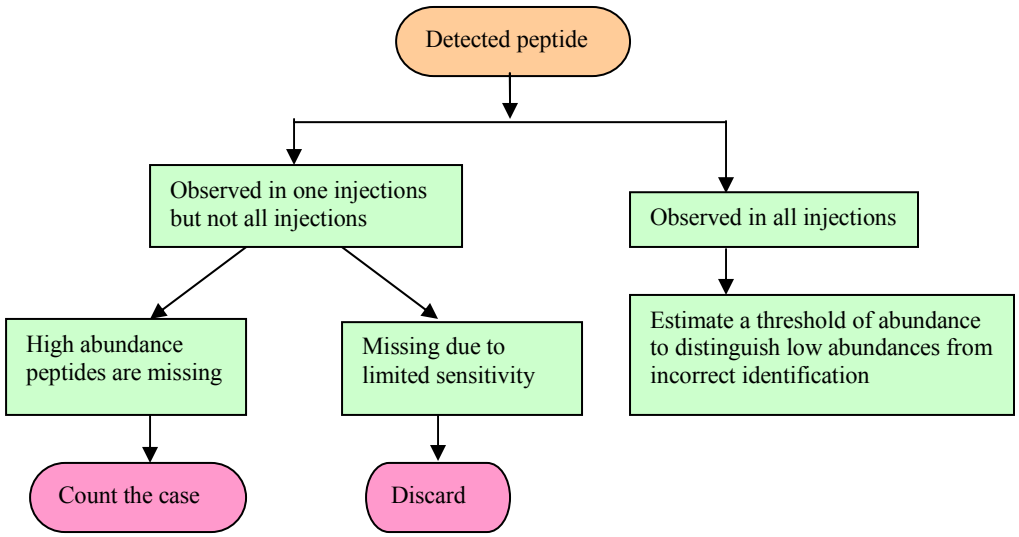

Figure 6.2: Exploring the missing data problem by evaluating different injection conditions.

6.3 Consistency test on MetaCore network analysis

We incorporate the software package MetaCore to explore biological interpretations of groups of proteins altered in abundance by the various treatments. MetaCore provides several graph-based tools to relate proteins altered in abundance with biological processes affected by treatments. One of them is the "analyze networks" tool illustrated in Figure 6.3. This tool displays both the proteins that we identify, called "targets", and the direction of protein abundance change, called "up- or down regulation" on signaling networks associated with the control of biological functions. Hence, the network becomes a directed graph with nodes that are proteins and edges that indicate how a biological process affects their abundance.

MetaCore uses an argument based on enrichment statistics to suggest which processes within its database of cellular signaling networks is the most likely explanation for the proteomic data. Enrichment statistics are based solely on the number of targets found on the network. The larger the number of targets the high the enrichment score and the more likely the signaling network is a valid interpretation of the proteomic data.

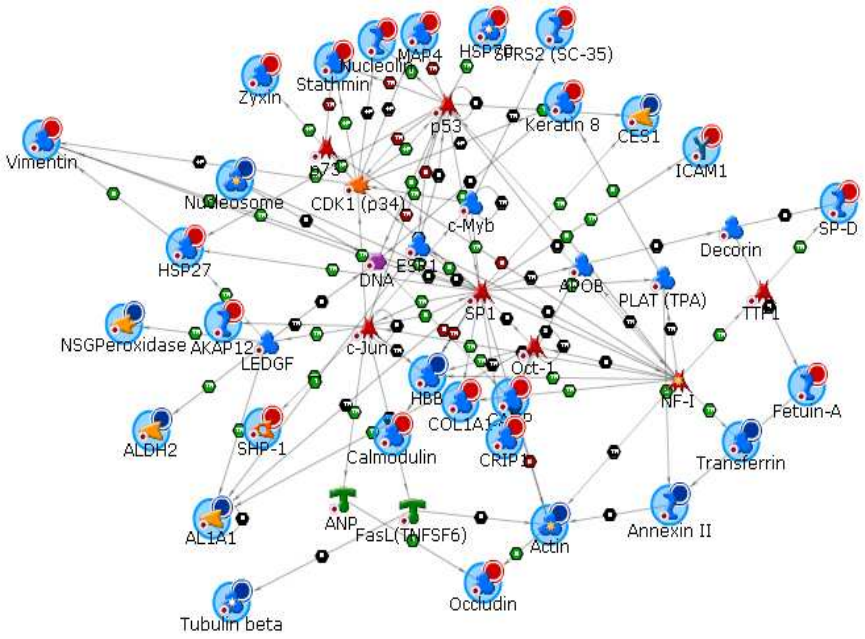

Figure 6.3: Sub-network associated with phagocytosis and apoptosis from application of analyze-network feature applied to proteins altered in abundance by all treatments.

The consistency of up and down regulation of targets with activation and inhibition within a network is not included in scores based solely on number of targets on a network. We propose that augmenting enrichment with consistency tests will increase our ability to discriminate between networks in the MetaCore database which are possible explanation for our proteomic observations.

Pulling information out of MetaCore's graphical expression of pathways is one of the difficulties associated with this project. Since MetaCore doesn't allow automated downloading of its pathway information due to its intellectual property, we have to start with manual data extraction from the graphic description on the pathway. We need to compose a property file, which contains nodes, edges, and their attributes of the MetaCore pathway, then reconstruct the defined graph by the nodes and edges data, and check the logic of the graph according to the attributes of the nodes and edges. For example, if an edge denotes gene activation, the gene-product node should be up-regulated. We score each signaling network, which interprets our proteomic data, by the faction of consistent nodes and edges. Networks with highest consistency ratio are assigned greater confident in interpreting the biological processes implied by the proteomic data. Figure 6.4 shows a sample property file for a reconstructed graph.

We are also considering about replacement of MetaCore with other available pathway analysis tools. Draghici's [105] systems biology approach for pathway level analysis could be a good choice.

The ultimate goal for use of our dataflow and integrated data-analyzing tools is to generate reliable analyzed proteomics results for biological and disease level biomarker investigations. Clearly, such improvements will play an important role in making high-throughput, global quantitative proteomics an established part of biomedical research.

1	Node	Attribute		
2	APP	red		
3	Calreticulin	red		
4	Cathepsin D	red		
5	DJ-1	blue		
6	ENO1	blue		
7	PEA15	red		
8	RAI	red		
9	TMSB4X	blue		
10	p53	black		
11	PTEN	black		
12	AKT(PKB)	black		
13	c-Myc	black		
14	Caspase-8	black		
15				
16	From	To	Effect	Mechanism
17	RAI	p53	inhibition	-unspecified-
18	PTEN	p53	activation	binding
19	p53	PTEN	activation	transcription regulation
20	DJ-1	p53	activation	binding
21	p53	Cathepsin D	activation	transcription regulation
22	c-Myc	p53	activation	transcription regulation
23	DJ-1	PTEN	inhibition	-unspecified-
24	Cathepsin D	APP	inhibition	cleavage
25	Calreticulin	APP	activation	binding
26	DJ-1	AKT(PKB)	activation	-unspecified-
27	AKT(PKB)	PEA15	activation	???- phosphorylation
28	PEA15	Caspase-8	inhibition	binding
29	ENO1	c-Myc	inhibition	transcription regulation
30	c-Myc	TMSB4X	inhibition	transcription regulation

Figure 6.4: A sample property file to reconstruct a directed graph.

Bibliography

1. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, and Williams KL. Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev* **1996**, 13, 19-50.

2. Anderson NL, and Anderson NG. Proteome and proteomics: New technologies, new concepts, and new words. *Electrophoresis* **1998**, 19, 1853-1861.

3. Thongboonkerd V, Klein JB. Proteomics in Nephrology. Contrib Nephrol. *Basel, Karger* **2004**, 141, 1-10.

4. Aebersold R, and Mann M. Mass spectrometry-based proteomics. *Nature* **2003**, 422, 198-207.

5. Omenn GS, States DJ, Adamski M, and Blackwell TW, et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**, 5, 3226-3245.

6. Lin B, White JT, Lu W, and Xie T, et al. Evidence for the Presence of Disease-Perturbed Networks in Prostate Cancer Cells by Genomic and Proteomic Analyses: A Systems Approach to Disease. *Cancer Res* **2005**, 65, 3081-3091.

7. Qian WJ, Camp DG, and Smith RD. High-throughput proteomics using Fourier transform ion cyclotron resonance mass spectrometry. *Expert Rev. Proteomics* **2004**, 1, 87-95.

8. Granville CA, and Dennis PA. An Overview of Lung Cancer Genomics and Proteomics. *Am. J. Respir. Cell Mol. Biol.* **2005**, 32, 169-176.

9. Fang R, Elias DA, Monroe ME, and Shen Y, et al. Differential Label-free Quantitative Proteomic Analysis of Shewanella oneidensis Cultured under Aerobic and Suboxic Conditions by Accurate Mass and Time Tag Approach. *Mol. Cell Proteomics* **2006**, 5, 714-725.

10. Bondarenko PV, Chelius D, and Shaler TA. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. Anal. *Chem.* **2002**, 74, 4741-4749.

11. Wang W, Zhou H, Lin H, and Roy S, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **2003**, 75-4818-4826.

12. Zurich ETH. Algorithms for peptide identification by tandem mass spectrometry. Dissertation. **2006**.

13. Shen YF, Zhang R, Moore RJ, and Kim J, et al. Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. *Anal. Chem.* **2005**, 77, 3090-3100.

14. Wang LN. Artificial neural network prediction of peptide identification probabilities in LC-MS/MS based proteomics. Thesis. **2005**.

15. Perkins DN, Pappin DJC, Creasy DM, and Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, 20, 3551-3567.

16. Clauser KR, Baker P, and Burlingame AL. Role of Accurate Mass Measurement (+10 ppm) in protein identification strategies employing MS or MS/MS and Database searching. *Anal. Chem.* **1999**, 71, 2871-2882.

17. Eng JK, McCormack AL, and Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database JASMS. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 976-989.

18. Field HI, Fenyo D, and Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2002**, 2, 36-47.

19. Liu T, Qian WJ, Strittmatter EF, David GC, Gordon AA, Brian DT, and Richard DS. High-throughput comparative proteome analysis using a quantitative cysteinyl-peptide enrichment technology. *J. Proteome Res.* **2004**, 3, 760-9.

20. Nesvizhskii AI, Keller A, Kolker E, and Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, 75, 4646-4658.

21. Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, and Simpson RJ. CHOMPER: a bioinformatics tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. *Proteomics* **2002**, 2, 1097-1103.

22. Tabb DL, McDonald WH, Yates JR. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, 1, 21-26.

23. Han DK, Eng J, Zhou H, and Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol*. **2001**, 19, 946-951.

24. Moore RE, Young MK, and Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **2002**, 13, 378-386.

25. http://www.sbc.su.se/~pjk/kth-systems-biology-2003/whatis.html.

26. Hiroaki K. Systems biology: a brief overview. *Science* **2002**, 295, 1662-1664.

27. Draghici S, Khatri P, Tarca AL, and Amin K, et al. A systems biology approach for pathway level analysis. To be submitted.

28. Kanehisa M, and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **2000**, 28, 27-30.

29. Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Research* **2004**, 32, D277-D280.

30. Kanehisa M, Goto S, Kawashima S, and Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Research* **2002**, 30, 42-46.

31. Nakao M, Bono H, Kawashima S, and Kamiya T, et al. Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Informatics* **1999**, 10, 94-103.

32. Ekins S, Bugrim A, Brovold L, Kirillov E, and Nikolsky Y, et al. White paper: algorithms for network analysis.

33. http://www.pir.uniprot.org/.

34. http://david.abcc.ncifcrf.gov/.

35. Wang, W, Zhou H, Lin H, and Roy S, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **2003**, 75, 4818-4826.

36. Radulovic, D, Jelveh S, Ryu S, and Hamilton TG, et al. Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2004**, 10, 984-997.

37. Prakash A, Mallick P, Whiteaker J, and Zhang H, et al. Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* **2006**, 5, 423-432.

38. Listgarten, J, and Emili, A. Statistical and computational methods for comparative proteome profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, 4, 419-434.

39. Ahram M, Adkins JN, Auberry DL, Wunschel DS, and Springer DS. A proteomic approach to characterize protein shedding. *Proteomics* **2005**, 5, 123-131.

40. Ahram M, Strittmatter EF, Monroe ME, Adkins JN, Hunter JC, Miller JH, and Springer DL. Identification of shed proteins from Chinese hamster ovary cells: application of statistical confidence using human and mouse protein database. Proteomics **2005**, 5, 1815-1826.

41. Hanash S. Disease proteomics. *Nature* **2003**, 422, 26-232.

42. Kearney P, and Thibault P. Bioinformatics meets proteomics – bridging the gap between mass spectrometry data analysis and cell biology. J. *Bioinform. Comp. Biol*. **2003**, 1, 183-200.

43. Tyers M, and Mann, M. From genomics to proteomics. *Nat. Rev. Genet.* **2003**, 422, 193-197.

44. Lilien RH, Farid H, and Donald BR. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Comput. Biol.* **2003**, 10, 925-946.

45. Strittmatter EF, Kangas LJ, Petritis K, and Mottaz HM, Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* **2004**, 3, 760-769.

46. http://www.ama-assn.org/ama/pub/category/3668.html.

47. Sadygov RG, Eng J, Durr E, and Saraf A III, et al. Code developments to improve the efficiency of automated MS/MS spectra interpretation. *J.Proteome.Res.* **2002**, 1, 211-215.

48. Ahram M, Adkins, JN, Auberry DL, Wunschel DS, and Springer DL. A proteomic approach to characterize protein shedding. *Proteomics* **2005**, 5, 123-131.

49. Smith RD, Anderson GA, Lipton MS, and Pasa-Tolic L, et al. An accurate mass tag strategy for quantitative and high-throughput. proteome measurements. *Proteomics* **2002**, 2, 513-523.

50. Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolic L, and Smith RD. Utility of accurate mass tags for proteome-wide protein identification. *Anal.Chem.* **2000**, 72, 3349-3354.

51. Pasa-Tolic L, Masselon C, Barry RC, Shen Y, and Smith RD. Proteomic Analyses using an Accurate Mass and Time Tag Strategy. *Biotechniques* **2004**, 37, 621-33, 636.

52. Matt Monroe and Kyle Littlefield, PNNL, 2004.

53. Nesvizhskii AI, Keller A, Kolker K, and Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, 75, 4646-4658.

54. Nesvizhskii AI, and Aebersold R. Interpretation of shortgun proteomic data. *Mol. Cell. Proteomics* **2005**, 4, 1419-1440.

55. Yu CG, Lin Y, Sun SW, Cai JJ, Zhang JF, Zhang Z, Chen RS, and Cheriton DR. An iterative algorithm to quantify the factors influencing peptide fragmentation for MS/MS spectrum. *Proceedings of the Conference CSB* **2006**.

56. Searle SR, Linear models for unbalanced data. *John Wiley & Sons*, New York, **1987**.

57. Pinheiro JC, and Bates DM. Mixed effects models in S and S-Plus. Springer Verlag, New York **2000**.

58. Daly and Anderson, manuscript in preparation.

59. Sharpe and Anderson, manuscript in preparation.

60. Cosio PMG, and Cosio MG. Disease of the airways in chronic obstructive pulmonary disease. *Eur. Respir. J. Suppl.* **2001**, 34, 41s-49s.

61. Cosio MG, Majo J, and Cosio MG. Inflammation of the airways and lung parenchyma in COPD: role of T cells. *Chest* **2002**, 121, 160S-165S.

62. Feghali CA, and Wright TM. Cytokines in acute and chronic inflammation. *Front Biosci.* **1997**, 2, d12-d26.

63. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, and Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am. J. Respir. Crit Care Med.* **2001**, 163, 1256-1276.

64. Barnes PJ, Chowdhury B, Kharitonov SA, and Magnussen H, et al. Pulmonary biomarkers in chronic obstructive pulmonary disease. *Am. J. Respir. Crit Care Med.* **2006**, 174, 6-14.

65. Brusselle GG, Bracke KR, Maes T, and D'hulst AI, et al. Murine models of COPD. *Pulm. Pharmacol. Ther.* **2006**, 19, 155-165.

66. Cavagna G, Foa V, and Vigliani EC. Effects in man and rabbits of inhalation of cotton dust or extracts and purified endotoxins. *Br. J. Ind. Med.* **1969**, 26, 314-321.

67. Michel O, Duchateau J, and Sergysels R. Effect of inhaled endotoxin on bronchial reactivity in asthmatic and normal subjects. *J. Appl. Physiol* **1989**, 66, 1059-1064.

68. Schwartz DA, Thorne PS, Yagla SJ, and Burmeister LF, et al. The role of endotoxin in grain dust-induced lung disease. *Am. J. Respir. Crit Care Med.* **1995**, 152, 603-608.

69. Smid T, Heederik D, Houba R, and Quanjer PH. Dust- and endotoxin-related acute lung function changes and work-related symptoms in workers in the animal feed industry. *Am. J. Ind. Med.* **1994**, 25, 877-888.

70. Vogelzang PF, van der Gulden, JW, Folgering H, and Kolk JJ, et al. Endotoxin exposure as a major determinant of lung function decline in pig farmers. *Am. J. Respir. Crit Care Med*. **1998**, 157, 15-18.

71. Nevalainen M, Raulo SM, Brazil TJ, and Pirie RS, et al. Inhalation of organic dusts and lipopolysaccharide increases gelatinolytic matrix metalloproteinases (MMPs) in the lungs of heaves horses. *Equine Vet. J.* **2002**, 34, 150-155.

72. Spond J, Billah MM, Chapman RW, and Egan RW, et al. The role of neutrophils in LPS-induced changes in pulmonary function in conscious rats. *Pulm. Pharmacol. Ther.* **2004**, 17, 133-140.

73. Toward TJ, and Broadley KJ. Goblet cell hyperplasia, airway function, and leukocyte infiltration after chronic lipopolysaccharide exposure in conscious Guinea pigs: effects of rolipram and dexamethasone. *J. Pharmacol. Exp. Ther.* **2002**, 302, 814-821.

74. Vernooy JH, Dentener MA, van Suylen RJ, Buurman WA, and Wouters EF. Long-term intratracheal lipopolysaccharide exposure in mice results in chronic lung inflammation and persistent pathology. *Am. J. Respir. Cell Mol. Biol.* **2002**, 26, 152-159.

75. Vernooy JH, Dentener MA, van Suylen RJ, Buurman WA, and Wouters EF. Intratracheal instillation of lipopolysaccharide in mice induces apoptosis in bronchial epithelial cells: no role for tumor necrosis factor-alpha and infiltrating neutrophils. *Am. J. Respir. Cell Mol. Biol.* **2001**, 24, 569-576.

76. Vlahos R, Bozinovski S, Gualano RC, Ernst M, and Anderson GP. Modelling COPD in mice. *Pulm. Pharmacol. Ther.* **2006**, 19, 12-17.

77. Lee KM, Renne RA, Harbo SJ, and Blessing JC, et al. Pulmonary response of AKR/J mice exposed to LPS and/or cigarette smoke for up to 3-weeks via nose only inhalation. *The Toxicologist* **2006**, 90, 340.

78. Lee KM, Renne RA, Harbo SJ, and Blessing JC, et al. 3-week inhalation exposure to cigarette smoke and/or lipopolysaccharide in AKR/J mice. *Inhalation Toxicology* **2007**, 19, 23-35.

79. Bogdanov B, and Smith RD. Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.* **2005**, 24, 168-200.

80. Purvine S, Picone AF, and Kolker E. Standard mixtures for proteome studies. *OMICS.* **2004**, 8, 79-92.

81. Daly DS, Anderson KK, Panisko EA, and Purvine SO, et al. submitted to Journal of Proteome Research.

82. McCullagh P, and Nelder JA. Generalized linear models. *Chapman & Hall/CRC*, Boca Raton **1989**.

83. Sharp JL, Anderson KK, Daly DS, and Auberry DL, et al. Inferring protein associations using protein pull-down assays *Proceedings of the American Statistical Association*, Biometrics Section, Alexandria, VA **2006**.

84. Meng QR, Gideon KM, Harbo SJ, and Renee RA, et al. *Inhal. Toxicol.* **2006**, 18, 555-568.

85. Janardhan KS, Appleyard GD, and Singh B. *Histochem. Cell Biol.* **2004**, 121, 383-390.

86. Mothersill C, and Seymour CB. Radiation-induced bystander effects - implications for cancer. *Nat. Rev. Cancer* **2004**, 4, 158-164.

87. Morgan WF. Non-targeted and delayed effects of exposure to ionizing radiation: I. Radiation-induced genomic instability and bystander effects in vitro. *Radiat. Res.* **2003**, 159, 567-580.

88. Morgan WF, Non-targeted and delayed effects of exposure to ionizing radiation: II. Radiation-induced genomic instability and bystander effects in vivo, clastogenic factors and transgenerational effects. *Radiat. Res.* **2003***,* 159, 581-596.

89. Ullrich RL, and Ponnaiya B. Radiation-induced instability and its relation to radiation carcinogenesis. *Int. J. Radiat. Res.* **1998***,* 74, 747-754.

90. Ahram A, Strittmatter EF, Monroe ME, Adkins JN, Hunter JC, Miller JH, and Springer DL. Identificaion of shed proteins from CHO cells: application of statisticl confidence using human and mouse protein databases. *Proteomics* **2005**, 5, 1815-1826.

91. Springer DL, Ahram M, Adkins JN, Kathmann LE, and Miller JH. Characterization of media conditioned by irradiated cells using proteome-wide high-throughput mass spectrometry. *Radiat. Res.* **2005**, 164, 651-654.

92. European Bioinformatics Institute ftp://ftp.ebi.ac.uk/pub/database/IPI/curent/.

93. University of Maryland, Human Protein Database.

94. Habermann B, Oegema J, Sunyaev S, and Shevchenko A. The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol. Cell. Proteomics.* **2004**, 3, 238-249.

95. HyperStat online contents, http://www/davidmlane.com/hyperstat/B112114.html.

96. Kim GJ, Fiskum GM, and Morgan WF. A role for mitochondrial dysfunction in perpetuating radiation-induced genomic instability. *Cancer Res.* **2006**, 66, 10377-10383.

97. Butow RA, and Avadhani NG. Mitochondrial signaling: the retrograde response. Mol. Cell 2004, 14, 1-15.

98. Rossignol R, Faustin B, Rocher C, Malgat M, Mazat JP, and Letellier T. Mitochondrial threshold effects. *Biochem J.* **2003**, 370, 751-762.

99. Young HD. Statistical treatment of experimental data. *McGraw-Hill*, New York, **1962**.

100. http://www.itl.nist.gov/div898/handbook/prc/section4/prc43.htm.

101. Eng JK, McCormack AL, and Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 976-989.

102. Lee KM, Renne RA, Harbo SJ, and Blessing JC, et al. Pulmonary response of AKR/J mice exposed to LPS and/or cigarette smoke for up to 3-weeks via nose only inhalation. *The Toxicologist* **2006**, 90, 340.

103. Lee KM, Renne RA, Harbo SJ, and Blessing JC, et al. 3-week inhalation exposure to cigarette smoke and/or lipopolysaccharide in AKR/J mice. *Inhalation Toxicology* **2007**, 19, 23-35.

104. Fang R, Elias DA, Monroe ME, and Shen Y, et al. Differential label-free quantitative proteomic analysis of shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol. Cell Proteomics* **2006**, 5, 714-725.

105. Draghici S, Khatri P, Tarca AL, Amin K, and Done A, et al. A systems biology approach for pathway level analysis. (A work to be submitted)