Mining deeper into the proteome: computational strategies for improving depth and breadth of coverage in high-throughput protein identification studies

by

Peter J. Ulintz

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2008

Doctoral Committee:

Professor Philip C Andrews, Chair
Professor Daniel M Burns Jr
Associate Professor Janine R Maddock
Associate Professor Rory Marcus Marks
Assistant Professor Alexey Nesvizhskii

Acknowledgements

Project-specific Acknowledgements

Chapter 2: Machine Learning: The work in this chapter was performed in collaboration with Ji Zhu from the Department of Statistics, and Zhaohui S. Qin from the Department of Biostatistics at the University of Michigan. The project evolved out of a lively term project for Steve Qin's Biostatistics class, and I am very grateful to both collaborators for their patience and effort in teaching me fundamental statistical learning approaches. We originally dedicated this work to the memory of Leo Breiman for his outstanding contributions to the study of statistics and machine learning. I want to also thank all members of the National Resource for Proteomics and Pathways who contributed to this work: namely Angela Walker for TOF/TOF mass spectrometry, Donna Veine and John Strahler for sample processing of the purified Genway proteins, and Tom Blackwell and Jayson Falkner for invaluable feedback on the manuscript. Thank to Andy Keller and Alexey Nesvizhskii for making the SEQUEST data available, and for the open release of the PeptideProphet software. Thanks also to Daniela Eggle for her contribution to the original support vector machine testing of the SEQUEST dataset. Lastly, one very thorough reviewer for the submitted manuscript of this work was very insightful and helpful, and the work was significantly improved by their feedback—you know who you are! This work was supported in part by the National Resource for Proteomics and Pathways funded by NCRR (P41 RR 18627-01 to P.C.A).

Chapter 3 and 4: MS^2/MS^3 and MSA: The work in these chapters was done in collaboration with Bernd Bodenmiller and Ruedi Aebersold, from the Institute of Molecular Systems Biology, in the Swiss Federal Institute of Technology in Zurich. All credit for the beautiful phosphopeptide data is theirs. Thanks to Alexey Nesvizhskii for his patient mentoring, particularly with the need to fill in my knowledge gaps of his prior work, and in dealing with more rounds of debugging than necessary. Thanks to Anastasia Yocum for her contributions to the effort, for the benefit of her mass spec knowledge, and for debating me on controversial issues. This work was supported in part by NIH/NCI Grant CA-126239 to AIN, NIH/NCRR - National Resource for Proteomics and Pathways Grant #P41-18627 to PCA, and with funds from NIH/NHLBI under contract No. N01-HV-28179 to RA. Bernd Bodenmiller is the recipient of a fellowship

by the Boehringer Ingelheim Fonds. I want to acknowledge Steven Tanner and the UCSD Computational Research Group for the free availability of their code; the InsPecT python library is a pleasure to use! All annotated spectra in this manuscript were generating using the Label.py and MakeImage.py modules available in the InsPecT library.

Chapter 5: Dengue. Many individuals have contributed to the generation of dengue data over the past several years. Primarily, Rory Marks must be thanked for providing the sample, but also for his mentorship and good humor. The intact-iTRAQ data was generated by Matt Willetts Marjorie Minkoff at ABI. The list of NRPP members past and present who have contributed dengue data are (in order of appearence): Donna Veine, Eric Simon, Sarah Volk, Antonia Chan, Pratik Jagtap, Anastasia Yocum, Laurette Prely, John Strahler, and Angela Walker.

Chapter 6: Additional work. The Zymogen granule project belongs to Xuequn Chen, and all creative credit for this project belongs to him. Thanks to Eric Simon for the continuing partnership on the phosphopetide isoelectric focusing project. Credit for the idea of using pI to enrich for methylesterified phosphopeptides goes to Phil Andrews, who thought of it years ago.

Personal Acknowledgements

It is with great honor that I acknowledge my teachers. I first thank my Mom for keeping me not only in school but in better schools (sometimes illegally!), so that I had the opportunity to make something of myself. My wife perhaps is owed the greatest acknowledgement for her encouragement and patience in this process; without her all this would be much more meaningless. I thank my mentor Phil Andrews for providing a great environment for exploration and growth since 1997; primary thanks goes to him for most of my knowledge of proteomics, and for introducing me to many other people who have had a significant impact on my scientific view and career. Janine Maddock provided a great deal of support both early and late in this process, thanks for the trip to TIGR and the excuse to use MAGPIE. Thanks to Dan Burns for teaching the first informatics course I formally took, for years of encouragement, and for many precise and hilarious analogies. Alexey Nesvizhskii has provided fantastic mentorship over the past couple years; his work is perhaps the most obvious overall influence on the approaches described in this manuscript. I'm grateful to Rory Marks for his patience in explaining biological aspects of some of this work, and for providing a great overall perspective on work and academics.

Earlier teachers deserve acknowledgement for their influence on my academic career; I would not be pursuing science today without their help. Thanks to Robert Slocum for making modern physics fascinating, and for chess games, cigarettes and scotch in his office (you can't do that on campus anymore!). Thanks to Paul Keyes for introducing me to the fascinating world of statistical physics, for mentoring me in my undergraduate days, and for a great day sailing. Luc Wille was an incredible mentor at Florida Atlantic University, providing great flexibility, guidance and humor during a difficult time.

I'm grateful to all the fellow students and post-docs I've had the pleasure to work with. Particular thanks go to Takis Papoulias for never sacrificing theory for practicality; Jayson Falkner for keeping me honest; Eric Simon for many many lunch discussions of science, politics, and morality; and Anastasia and Todd Yocum for being my best Ann Arbor friends during the last year or so of graduate school.

Lastly, I'm incredibly grateful for the opportunity to study the Tibetan language while in graduate school from Gareth Sparham. I'm also deeply honored to acknowledge my other teachers who so immensely enriched my mental atmosphere: Geshe Michael Roach, Christie McNally, Ven Lobsang Chukyi, and the rest of the World-View community.
བདག་གིས་གྲུབ་པས་ཕྱག་འཚལ་ལོ།

Table of Contents

List of Figures

List of Tables

ABSTRACT

Mining deeper into the proteome: computational strategies for improving depth and breadth of coverage in high-throughput protein identification studies

by

Peter J. Ulintz

Chair: Philip C. Andrews

The proteomics field is driven by the need to develop increasingly high-throughput methods for the identification and characterization of proteins. The overall goal of this research is to improve the success rate of modern high-throughput proteomics studies. The focus is on developing computational strategies for increasing the number of identifications as well as improving the ability to distinguish new forms of proteins and peptides. Several studies are presented, addressing different points in the proteomics analysis pipeline. At the most fundamental data analysis level, methods for using modern machine learning algorithms to improve the ability to distinguish correct from incorrect peptide identifications are presented. These techniques have the potential to minimize the need for manual curation of results, providing a significant increase in throughput in addition to increased identification confidence.

Non-standard types of mass spectrometry data are being generated in specific contexts. Specifically, phosphoproteomics often involves the generation of MS^3 spectra.

These spectra alleviate problems associated with MS^2 fragmentation of phosphopeptides, but utilizing the additional information contained in these spectra requires novel informatics. Several strategies for accommodating this additional information are presented. A statistical model is developed for translating the information contained in the coupling of consecutive MS^2 and MS^3 spectra into a more accurate peptide identification probability score. Also, methods for combining MS^2 and MS^3 data are explored.

A newer mass spectrometry methodology useful for phosphoproteomics has recently been introduced as well, termed multistage activation (MSA). A comparative study of this and other methods is presented aimed at determining an optimal method for generating phosphopeptide identifications, focusing not only on data analysis techniques, but also on the mass spectrometry methodologies themselves.

A dataset is presented from a differential study of a human cell line infected with the dengue virus. The study explores the complementarity of different fractionation methods in generating more unique protein identifications. A discussion of a statistical mixture model that utilizes relative quantification information to classify identified peptides into two categories based on their membrane topology is given in the final chapter. Finally, a comment on utilizing pI information to enrich for phosphopeptides is provided.

Chapter 1

Introduction

1.1 A brief history

History may very well remember the past fifty years, and the dawn of the 21st century, as being primarily characterized by two major technological revolutions: those of computers and life science. The discovery of the structure of DNA in 1953, and the subsequent understanding that DNA nucleotide sequence completely specifies the amino acid sequence of proteins, laid the foundation for the modern understanding of how living systems function. Concordantly, IBM shipped its first stored-program computer, the 701, in 1953, and the first FORTRAN program was successfully run the following year. By the time the genetic code was elucidated in the early- to mid-sixties, IBM's annual sales revenue for computer-based products exceeded $1 billion dollars.[1] Microsoft was founded in 1975, the year Sanger and his colleagues developed their DNA sequencing methods. The year 1983, marked by the advent of the polymerase chain reaction (PCR) technique for amplification of DNA that brought genetic studies to the benches of so many more scientists, was matched by computers achieving the one billion floating point computation per second benchmark, and the distribution of personal computers to the mass market was well underway.

[1] And the first Computer Science Ph.D. was granted in 1965 at the University of Pennsylvania.

The genetic code not only provided a conceptual framework for understanding the function and evolution of living organisms, but provided a basis for representing the information content of a biological entity as essentially text: a linear string of a small alphabet of symbols. As soon as protein and nucleotide sequences started being known, methods for comparing these sequences to derive their evolutionary history started to be developed. In addition to constructing the first phylogenetic model of a protein family, Margret Dayhoff formulated the first probabilistic distance matrices for comparing amino acid sequences, and published the *Atlas of Protein Sequences and Structure* in 1965 (1). The *Atlas* was one of the first collections of sequence information to be assembled, and evolved into to the first public repository of sequence information: the Protein Information Repository (PIR). 1981 marked the publication of the first sequence alignment algorithm (2), and the GenBank repository was founded the following year. Following the breakthrough of PCR and modern sequencing, vectors for large scale genomic sequencing were established by the late eighties. The first physical map of the genome of an organism (*E. coli*) was generated in 1987 (3), and the question of 'How much information is necessary to encode an organism?' began to be addressed. The disciplines of computer science and molecular biology thus became inextricably married, and the field of bioinformatics was born out of the necessity to manage, mine, and compute upon the large amount of biological information that was being generated at an increasing rate. One final landmark pivotal in the foundation of bioinformatics came in 1994, when Netscape Communications released the first version of the Navigator web browser, establishing the Web as the primary global information infrastructure necessary for publishing and sharing information. Almost a decade after that first physical map of *E. coli*, modern databases containing assemblies of entire genomes began to become available (4), culminating in the human genome assemblies in 2001 (5,6).

More directly related to the problem at hand, the late eighties also marked a major advance in the field of protein chemistry: the ability of proteins to be effectively ionized in an intact form without extensive fragmentation. The establishment of two ionization methods in particular, electrospray ionization (ESI) and matrix-assisted laser desorption and ionization (MALDI), heralded the beginning of large scale protein identification by

mass spectrometry. ESI quickly established itself as a popular technique due to its ability to be easily interfaced with liquid chromatography (LC) fractionation systems. Also, the fact that the technique resulted in multiply charged analytes allowed higher mass species to be detected in instruments with smaller mass ranges. In contrast, MALDI instruments typically had larger overall mass ranges, a feature of the relatively simple time-of-flight (TOF) mass analyzers most often coupled with the source. The characteristic of MALDI analytes to be singly charged also simplified the interpretation of the resulting spectra. Both types of instruments quickly became established as reliable means for identifying proteins and are still the dominant ionization techniques used for protein analysis. The importance of these techniques for protein elucidation is reflected by the awarding of the Nobel Prize to John Fenn for ESI and Koichi Tanaka for MALDI in 2002. With the establishment of these techniques, all the key components were in place for the foundation of proteomics research.

1.2 Proteomics, Past and Present

The term 'Proteomics' can be defined in many ways, but if one considers the term to apply narrowly to the ability to identify and quantify proteins in a high-throughput manner, then not only is it a means to quickly generate the necessary data but also a means to interpret it effectively. The field of proteomics was born not only from the ability to generate protein identification information in high-throughput provided by mass spectrometry, but also by the availability of public repositories of genome sequence information and a means to utilize this information for the interpretation of mass spectra. The third major component in the proteomics equation was that of robust protein separation technologies. In the early years, the term 'proteomics' was almost synonymous with two-dimensional gel electrophoresis (2DE), a separation method first developed in the mid-seventies (7, 8). The paradigm of focusing proteins from complex mixtures on 2D gels, excising the spots on those gels, digesting the proteins and analyzing the resulting peptides in a (MALDI-TOF) mass spectrometer became a default standard method for early proteomics research (9). In fact, it was at the first 2DE

meeting in Siena, Italy in 1994 that the term 'proteome' was first coined, by the Australian PhD student Marc Wilkins[2].

Since that time, proteomics has seen significant refinement in all three of these key areas. Perhaps the most significant shift, however, has occurred in the separations domain. Although 2DE is still extensively used, it has limitations. Not only is it labor-intensive (and thus time-consuming and costly, and more difficult to scale up), but it also can have reproducibility issues, and the fact that multiple proteins often fractionate in an individual spot on a gel creates interpretation and quantification problems. There are also solubility issues associated with gel-based approaches, and staining and extraction efficiencies can limit the sensitivity of the method. These reasons have resulted in a shift away from gel-based fractionation methodologies to other LC-based methods for groups requiring higher throughput. Early work by Donald Hunt's group (10), and subsequent automation and refinement by Yates et al. (11), demonstrated the power of the LC method in rapidly identifying a large number of proteins in a sample.

The digestion of complex protein extracts into even more complex peptide mixtures followed by the coupling of single- or multi-dimensional fractionation of complex protein mixtures to mass spectrometry (LC-MS/MS) has been termed "Shotgun Proteomics", coined in reference to the genomics analogy of assembling a large number of random pieces of overlapping DNA. The term "MudPIT" is often used in a synonymous manner, originally suggested by Yates et al. as an abbreviation of the term "multidimensional protein identification technology", and referring specifically to a tandem LC separation followed by mass spectrometry and database searching (12). This technology proved scalable and demonstrated the potential for a significant increase in identification rate over 2DE approaches, with the ability to identify hundreds or thousands of proteins in a single experiment (13-16).

Although shotgun proteomics analysis has now established itself as the *de facto* standard methodology for rapidly identifying proteins in complex mixtures, it is far from routine, and not without its own set of issues. The higher throughput of the technique has

[2] http://www.proteome.org.au/History-of-Proteomics/default.aspx

been won at an increase in the complexity of data analysis. Also, the difficulties of dealing with the very large number of components in a whole proteome were perhaps underestimated. Not only does this complexity create a resolution problem in fractionation space, but a dynamic range issue as well. Unlike the genomics area, in which DNA can be amplified many-fold for simplified analysis with technologies like PCR, proteomics must cope with naturally-occurring abundances of proteins in a sample. The proteins expressed in an organism, tissue, or cell at any given point in time vary dramatically in their levels of expression. The variation in expression of proteins in biological samples has been estimated as being as much as 10^6 orders of magnitude in yeast (17). In human serum, the most abundant protein (serum albumin) occurs at roughly 50 mg/ml. whereas the estimated level of lower copy number proteins such as transcription factors can be in the pg/ml range, a dynamic range of 10^{10}. This massive difference in expression creates a problem for any detection methodology.

Quantification, in fact, is one of the advantages the 2DE approach has over the LC-MS/MS approaches. Relative quantification information was available using image analysis on 2D gels. Mass spectrometry is not inherently quantitative, and the intensities of peaks in a spectrum are not necessarily an indication of the abundance of an analyte in a sample. Quantification information was thus lost in moving from 2DE to an LC-based fractionation approach. However, a number of techniques for deriving relative expression information in LC-MS/MS experiments have been developed. One such method is isotope-coded affinity tagging (ICAT), in which a labeled pair of reagents of differing isotopic mass are bound via an alkylating group to the cysteines of all proteins a population (18). The tag also contains a biotin group (in the earlier incarnation), permitting the targeted binding and selection of only the labeled peptides from the complex digested mixture. A similar and arguably preferable method to ICAT when applicable is the SILAC approach, in which isotopically-labeled amino acids are added to cell culture media deficient in those amino acids, growing the targeted cell system in the media. This latter method avoids chemical labeling or potentially noisy affinity purification steps, and provides a means of multiplexing beyond two labels by utilizing more than one amino acid. A newer methodology was developed by Darryl Pappin using

a set of stable isotope labeling reagents known as iTRAQ (19). Unlike the ICAT or SILAC reagents, iTRAQ labels are isobaric: the mass of any one labeled species is identical in the first dimension of mass spectrometry. Only when MS/MS fragmentation occurs on the isobarically labeled analyte mixture does the relative expression information become available. iTRAQ has the advantage of being a four-plex set of labels as originally developed, and an eight-plex set has become recently available (20). A large number of other stable isotope labeling methods have been developed for quantification in proteomics (21, 22) but for the sake of brevity are not discussed in detail here. Also, methods have been developed to quantify proteins without the use of labeling techniques, the so-called 'label free' approaches (23, 24). These methods exploit the fact that the total number of peptides that are identified for a protein correlates well with the protein abundance. In all, these quantification approaches allow LC-MS/MS approaches to be quantitative, and offer a more complete alternative to 2DE methodologies.

As mentioned, the ICAT labels have an additional advantage of allowing selective pull-down of the labeled peptides, greatly simplifying the overall protein mixture and therefore potentially providing for a deeper look into the proteome. More directed approaches which target specific subsets of the proteome have in general been increasingly adopted as ways of addressing the dynamic range issue. These approaches seek to utilize a discriminative feature of a set of proteins, such as the presence of a cysteine residue (18, 25), a glycosylation site (26), or a phosphate group (discussed below). Affinity tag- bait-based strategies have been particularly powerfully employed in high-throughput as a means to generate large and rich datasets for studying protein complexes, pathways, and protein-protein interactions (27, 28). More recently, targeted reaction monitoring methodologies are becoming more prevalent, in which the mass spectrometer is directed to look specifically for particular peptide masses or peptide/fragment-ion transitions in a biological sample (29, 30). This approach is quantitative when used in conjunction with isotopically labeled standard proteins. The methodology requires an advanced knowledge of specific peptide or protein forms, but allows a potentially dramatic increase in sensitivity, with the advantage of having much of the computational analysis done during assay development rather than being reliant on

large scale post-acquisition data analysis. Overall, directed studies focusing on specific subsets of proteins or peptides promise to allow a much greater penetration into the proteome.

1.3 Phosphoproteomics

With the possible exception of glycosylation, phosphorylation is the most highly-studied protein modification in proteomics research. Phosphorylation on serine, threonine, and tyrosine residues is generally held to be one of the most prevalent and biologically important post-translation modifications (PTMs). A primary role of phosphorylation is to act as a biological "on"/"off" switch for a protein activity or a cellular pathway in a specific and reversible manner (31, 32). The modification can alter the function of a protein in a number of ways: by affecting its activity, its stability, its subcellular localization, its ability to complex with other proteins, or by marking the protein for degradation (32). The fact that there are an estimated 518 protein kinases in humans (33) and 540 in mice (34), a number that is likely doubled in plants (35), and that at least one in three proteins are estimated to be phosphorylated at some point during their life cycle (32, 36), underscores the biological importance of this modification. Alterations in normal phosphorylation patterns have been implicated in a number of diseases, including cancer (37-39), diabetes (40), and Alzheimer's (41). The identification of phosphoproteins, and understanding the dynamics of this modification in response to cellular and environmental factors, is thus critical for elucidating the systems biology of complex disease mechanisms and global regulatory networks.

Prior to targeted proteomics technologies, detection of protein phosphorylation was somewhat low-throughput, conducted using approaches such as site-directed mutagenesis, Edman degradation, or two-dimensional thin layer chromatography and/or protein autoradiography of ^{32}P-labeled phosphoproteins (42). In fact, even with the use of mass spectrometry, it is only recently that larger numbers of phosphopeptides are beginning to be identified. A recent survey found that in 203 publications in which mass

spectrometry was used for phosphoprotein analysis between 1992 and 2003, only 1281 total phosphorylation sites were reported (42). Contrasted with more recent studies such as Li et al. which identified 2288 unique phosphorylation sites from 985 proteins in yeast (43), another yeast study by Donald Hunt's group identifying 1252 phosphorylation sites on 629 proteins (44), or a study by Molina et al. on human embryonic kidney 293T cells which produced 1435 unique phosphorylation site identifications from 500 proteins (45), it can be seen that the ability to rapidly identify many more phosphorylated protein forms is increasing rapidly.

There are several features of phosphoproteome analysis that needed to be overcome to achieve higher success rates in high-throughput studies. The primary difficulty is one that has already been mentioned: stoichiometry and dynamic range. Phosphoproteins are often expressed in relatively low amounts in a cell, and relatively few of these proteins exist in a phosphorylated form at any one time. Enrichment strategies are therefore necessary to identify the modified forms of these proteins in high-throughput. Initially, these strategies were sub-optimal, but more recent implementations are performing much better and are now at the point where an 80-90% enrichment can reliably be achieved (45, 46). There are several primary strategies for phosphopeptide enrichment: strong cation exchange (SCX), immobilized metal affinity chromatography (IMAC), titanium dioxide (TiO_2) and zirconium dioxide (ZnO_2) affinity, immunoaffinity, and chemical derivatization. The sensitivity of these strategies differs and depends on the specific method and the complexity of the sample. A recent comparison by Bodenmiller et al. of IMAC, TiO_2, and chemical derivatization using phosphoramidate chemistry found IMAC and the chemical derivatization methods to perform best, although all methods seemed to uniquely identify a large number of peptides suggesting that the techniques may be complementary (46).

A second major challenge for phosphoproteomics is that phosphopeptides can exhibit poor fragmentation in a mass spectrometer. This is due to the fact that the phosphate moiety is often the most labile element on the peptide. In the case of collision-induced dissociation (CID), much of the fragmentation energy used to produce a tandem (MS/MS or MS^2) mass spectrum often is absorbed in the dissociation of the phosphate

group. The resulting spectrum is often dominated by one or several peaks corresponding to the neutral loss of phosphoric acid, with little other fragmentation information useful for identification of the peptide sequence (47). This issue has been addressed using data-dependent MS^3 methodologies for generating mass spectra, typically on ion-trap instruments. Subjecting neutral-loss fragment masses to a further cycle of fragmentation often produces a spectrum with much more useful structural information on the peptide (48). Therefore, phosphopeptides have often been analyzed by automated data-dependent triggering of MS^3 acquisition whenever the neutral loss ion of the appropriate mass is detected in an MS^2 spectrum as a dominant peak (49-56). Although CID is by far the most commonly used methodology for peptide fragmentation, an alternative methodology is proving to be very useful for phosphoproteomics, electron transfer dissociation (ETD). This method functions by transferring electrons to the protonated peptide ions, with resulting fragmentation that is significantly different than CID. ETD spectra in general yield much more extensive backbone fragmentation than CID spectra, and have the property that labile PTMs are often preserved in the process (57). ETD has been demonstrated to be particularly effective for phosphopeptide analysis (44, 45).

A third source of difficulty for large-scale phosphoproteomics is one of data analysis. Informatics approaches for processing the results of phosphopeptide mass spectrometry data are not yet routine in many cases, specifically in the handling of non-standard types of spectra such as MS^3 or ETD spectra. As a result, manual curation and validation were often necessary for the spectra generated in phosphoproteomics studies, a significantly rate-limiting step. One of the major goals of this thesis is to address these particular issues.

1.4 Proteome Informatics

Informatics and protein mass spectrometry have been closely linked from the earliest days of proteomics. As soon as mass spectrometry began to be widely used for protein work, the need for automated software to assist in data analysis became apparent. Early algorithms had been developed for interpretation of the fast atom bombardment

(FAB) ionization spectra that preceded MALDI and ESI (58, 59). When these more successful ionization methods became available, other software quickly followed. By 1994, two classes of algorithms had already begun to be distinguished, corresponding to the types of spectra that these two ionization methods generated.

As mentioned earlier, initially MALDI was frequently matched with TOF mass analyzers, whereas ESI was most often paired with triple-quadrupole or ion-trap analyzers. In the MALDI-TOF approach, the paradigm for analyzing single protein isolates was to digest the protein with a proteolytic enzyme (typically trypsin) and measure the masses of the resulting fragments. The mass resolution of MALDI was sufficiently high to be able to uniquely identify a protein by the characteristic set of masses. This process came to be known as "peptide mass fingerprinting" (PMF). Algorithms were developed which took as input a "database" of known protein sequences and performed a theoretical digestion for each protein. Given an experimental mass spectrum, these algorithms would return the database proteins which generated a list of digestion masses best explaining the spectrum (60-64).

ESI spectra are fundamentally different than MALDI, in that the masses measured can be multiply charged. Moreover, the triple-quadrupole and ion-trap analyzers coupled with ESI sources have the capability of isolating specific fragment ions and subjecting them to further fragmentation (MS/MS, MS^2, or tandem spectra). Peptides fragment in predictable, sequence-dependent ways, and the resulting spectra can be interpreted to infer the amino acid sequence of the peptide. Algorithms that could match experimental MS/MS spectra to sequences in a database began to be published (65, 66).

Various components of these original scoring approaches for both types of spectra still exist, albeit enhanced in their current implementation: e.g. the MOlecular Weight SEarch (MOWSE) scoring model, the first to account for peptide length (62), was later developed into the Mascot algorithm (67). The original algorithm developed to perform an automated interpretation of MS/MS spectra, SEQUEST (65), is still arguably the best in its class, and represents the most successful single piece of software in proteome informatics. It functions by also performing a theoretical digestion of protein sequences

in a database, selecting a list of peptides with masses most similar to the parent mass that generated the MS/MS spectrum being analyzed. For each of these peptides, a theoretical fragmentation spectrum is generated and matched to the experimental spectrum. A set of scores is produced for each match, and the ranked list of matching peptide sequences is returned. SEQUEST was the first algorithm that was designed to take a large number of spectra as input and search them in an automated manner. The analysis paradigm that was established with the SEQUEST algorithm has been immensely successful, and is as fundamental to MS-based proteomics research as BLAST is to sequence alignment in the genomics field.

As the Web began to be established, various incarnations of MS interpretation software began to be made available via a web server. As the nineties progressed, the MOWSE functionality was rewritten to include MS^2 searching and was commercialized as Mascot. SEQUEST had been commercialized early on, controversially in that intellectual property rights had been established around the idea of MS^2 database searching, limiting the ability of other groups to develop similar tools. A number of other groups did ultimately develop significant tools, however, for both MS^2 and PMF spectra, including ProteinProspector (68), Profound (69), and PeptIdent/MultiIdent (70). The fundamentals of database searching strategies and scoring had been established, and attention began to shift to automation and throughput. Instrumentation had developed to the point that an increasing number of spectra were being generated, and researchers required more automated ways of searching collections of spectra rather than searching on a spectrum-by-spectrum basis. The software packages that were able to do this successfully, permitting local installations that could be scripted or take batch input, survived.

Another class of algorithm had begun to be defined as well, the so-called de novo algorithms for interpretation of MS^2 data (71-74). The goal of these algorithms is to infer the sequence of a peptide directly from the spectrum without the aid of a sequence database. Once the sequence is extracted, standard sequence alignment algorithms could be used to obtain the identity of the originating protein. De novo interpretation of an MS/MS spectrum is a challenging problem, however, albeit a seductive one from an

informatics perspective. A number of elegant algorithms have developed for de novo sequence analysis, but have never been able to achieve the success rate in a production environment that their database search equivalents have for known protein sequences. Nevertheless, some of the techniques used in these approaches have found their way into more standard approaches, and have led to the development of "hybrid" strategies for mass spectrometry data interpretation, as well as approaches for aligning and clustering mass spectra (75, 76).

By the turn of the century, PMF had declined as an overall approach in relationship to MS^2 approaches, since a tryptic mass fingerprint did not typically contain sufficient information to uniquely identify a protein, particularly for complex genomes and also as the sizes of the commonly-used public sequence repositories got larger. A notable exception to this fact is the use of the accurate mass tagging (AMT) approach pioneered by Dick Smith, which takes advantage of the high mass accuracy available on high-end Fourier Transform (FT) instruments as well as other information such as retention time on a column to uniquely identify peptides (77, 78). Instrumentation had continued to improve the rate at which spectra were being generated as well. The rate-limiting step in high-throughput proteomics projects was no longer the rate at which spectra could be acquired, but the rate at which it could be interpreted. The database searching strategies would almost always produce a set of "hits", a list of matching peptides, for each spectrum. The issue became one of whether these top scoring peptide assignments were correct or not. Straightforward thresholding methods for selecting correct hits above a specified cutoff score were simply not performing as adequately as desired for most researchers: many correct hits were falling below threshold values and being lost, while many incorrect assignments were being retained. Also, the heterogeneity of the various search algorithm scoring methods made comparison of results from different approaches more difficult, and biological researchers had trouble knowing which scoring thresholds could be relied upon as a discriminator for a confident assignment. Methods to put search results on a more firm statistical basis began to be developed. Intuitively, biologists are familiar with measures such as a probability score

or an expectation value; such scores needed to be defined for the proteomics domain as well.

The statistical approaches for data validation tend to follow two general strategies, termed single-spectrum and global strategies (79). Single-spectrum results compute an expectation value from the distribution of all scores produced by searching that spectrum against a database (80-83); this approach is exemplified by the X!Tandem and OMMSA search engines. Global strategies model the distribution of (typically) top-scoring hits from all spectra in the dataset rather than scoring search results individually. This approach is exemplified by the PeptideProphet and decoy database methods (84-87). PeptideProphet is particularly relevant to the work discussed here. It functions by fitting "correct" and "incorrect" distributions to the overall distribution of database search scores, calculating a Bayesian probability score based on these distributions. The Bayesian scoring framework is such that additional information that may be useful for discrimination of correct from incorrect peptide identifications can be modeled and accommodated into the score. The complementary tool to PeptideProphet, ProteinProphet, implements this scoring framework by utilizing peptide probabilities and other discriminatory information (such as the 'number of sibling peptides', NSP) in the calculation a probability for a protein identification (88). Recent work combining the decoy-database strategy with PeptideProphet has been published, and should provide a significant performance increase (85). Also, the global and single-spectrum approaches can in fact be complementary, and utilized within the same framework (79).

Recent years have seen an explosion in the number of proteome informatics publications. A number of new approaches to interpreting large-scale data have been developed, such as spectral library searching (89-91) and spectral clustering (92-95), with exciting results. Blind searching (96, 97) and top-down strategies (98, 99) show promise for expanding the typical search space to accommodate new or rare PTMs. Platforms such as Tranche[3] that permit simplified data management, integration, storage and dissemination, and provide a framework for mining very large collections of data from

[3] http://tranche.proteomecommons.org

multiple sources, promise to provide new insights and dramatically improve the accessibility of both data and software tools to many more users. Finally, the development of standardized data representations and openly available code libraries (100-102) should facilitate the development of increasingly advanced applications for protein and peptide identification.

1.5 Overview of Projects

Although there are a large number of disparate technologies that can be classified under the domain of proteomics, most work in the area follows a somewhat standard protocol. The method of high-throughput protein identification using LC-MS/MS can be defined by a number of well-defined steps, exemplified by the workflow in Figure 1-1. Proteins must be extracted from a biological sample and prepared for analysis by removing unwanted contaminants, or treating with inhibitors or phosphatases, etc., depending on the requirements of the experiment. The proteins are then digested with a site-specific protease to make the mixture components more amenable to mass spectrometry, and also more soluble. An enrichment step may occur: ion metal affinity chromatography (IMAC) or strong cation exchange (SCX) chromatography, for example, may be used to enrich phosphopeptides from non-phosphopeptides. The resulting mixture of peptides is then resolved into simpler components, typically by one dimension or two dimensions of fractionation. The final dimension is most typically reversed-phase high-performance liquid chromatography (RP-HPLC), separating the component peptides based on their hydrophobicity. The RP-HPLC column is often connected directly online to the mass spectrometer, such that the separated peptides eluting off the column are measured directly by the instrument. The instrument generates a collection of spectra for the experiment, which must be processed and analyzed. There is a computational pipeline associated with the final data analysis procedure that is not outlined, but involves extracting the spectra from the instrument, selecting a list of peaks for each spectrum, searching the spectum using a database search platform such as SEQUEST, and processing the results from the search using tools such as Peptide/ProteinProphet.

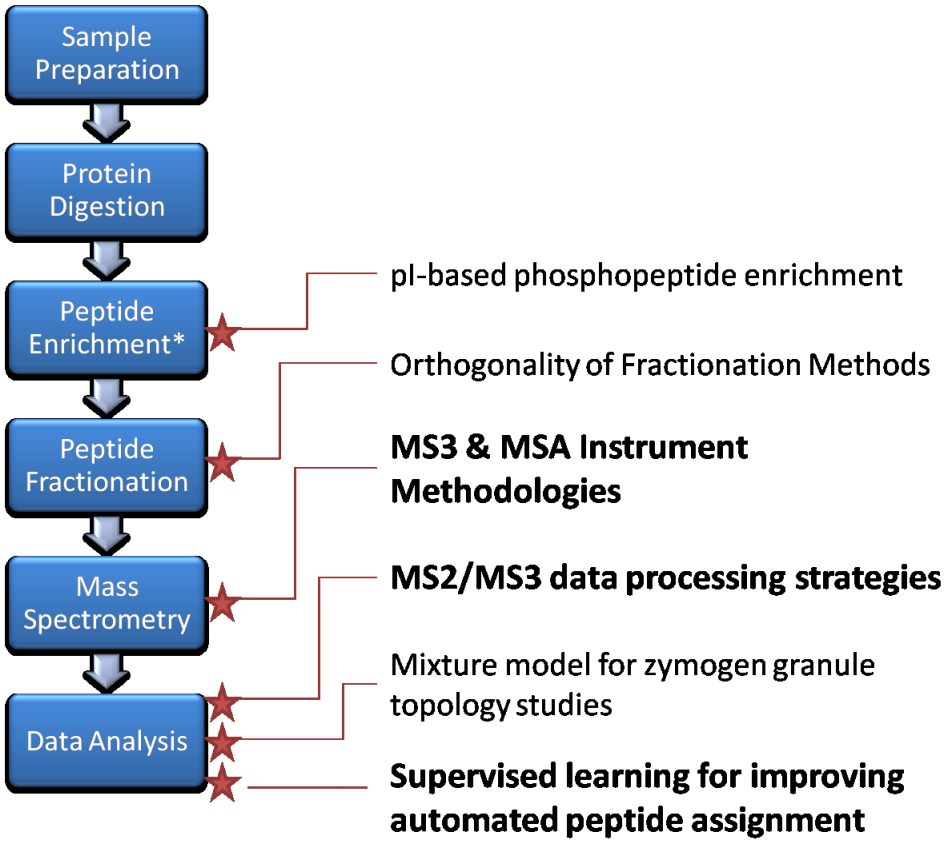


Figure 1-1. Outline of a standard LC-MS/MS proteomics analysis methodology. Protein samples are digested with a proteolytic enzyme, typically trypsin. The resulting peptide mixture may be optionally (indicated with a '*') enriched for a particular class of peptide, e.g. phosphopeptides. The mixture is then separated into simpler fractions using one or more dimensions of separation. Note that enrichment and fractionation steps can be interchanged, or even precede digestion. The resulting fractionated mixture is then introduced into the mass spectrometer, generating a collection of spectra that must be interpreted in an automated way. The location of particular research topics that are explored are illustrated in the schematic. The projects listed in bold are given particular attention, corresponding to chapters in this thesis.

The overall goal of the work presented in this thesis is to utilize informatics strategies to improve upon the proteomics production pipeline with the intent of enhancing the depth and breadth of proteome coverage. Informatics tools were used to both evaluate efficacy of specific strategies and technologies in this regard and also suggest additional strategies. Moreover, new informatics approaches were developed to improve the quality and extent of information available from existing technologies.

Outlined on the figure are the set of individual studies that were conducted, diagramming the specific areas being addressed in the proteomic production 'pipeline'. The studies indicated in bold are the ones that will be discussed in detail and form the bulk of the research effort. The other studies will be touched upon in less detail.

The research discussed in Chapter Two of this thesis addresses the area of statistical learning briefly introduced in section 1.4. As mentioned, the PeptideProphet publication was the first to develop the notion of global statistical analysis of scoring, implementing a model which fits "correct" and "incorrect" distributions to the distribution of scores and utilizing these distributions to derive a probability score (78). The work outlined in Chapter Two addresses a similar issue but in a different manner. The problem of whether a hit is correct or incorrect is formulated as a strict classification problem, a well-defined area of machine learning. Using various data attributes produced by a search algorithm, the work evaluates the performance of standardized and optimal pattern classification algorithms in classifying mass spectrometry database search results. The work explores these methods as flexible frameworks that are somewhat independent of not only the search engine used but also the specific attributes and scores produced by the individual search engines. Moreover, the methods are able to provide data on which attributes produced by a scoring approach are most discriminative in selecting a correct hit from an incorrect hit. Results are presented on a standardized dataset of SEQUEST results generated on an ESI ion-trap instrument, as well as a set of standardized results generated in our research lab on a MALDI-TOF/TOF instrument searched using SpectrumMill, a current implementation of the ProteinProspector suite of programs further developed by Karl Clauser at Millennium Pharmaceuticals and commercialized by Agilent Technologies[4].

There is a heavy focus in this research on phosphoproteomics, which begins to be addressed directly in Chapter Three. This is primarily for the reasons indicated above. LC-MS/MS phosphopeptide analysis methodologies most typically generate MS^3 spectra, yet these spectra are not optimally utilized by current data processing methods. The work

[4] http://www.chem.agilent.com/scripts/pds.asp?lpage=7771

in Chapter Three outlines a computational framework for processing MS^3 data in conjunction with MS^2 data, making efficient use of the additional information available from linking MS^2 and MS^3 scans. This information is translated into an adjustment to the probability score for a peptide identification. The methodology defined need not be solely used for phosphopeptide data however; MS^3 data is often generated in other contexts as well. The methods developed are demonstrated first on a digestion of known standard proteins (not-phosphorylated), then demonstrated on a complex phosphopeptide mixture. Overall, the resulting probabilities demonstrate an improvement in the ability to discriminate correct identifications from incorrect, a particularly important problem given the difficulty of identifying phosphopeptides.

Another topic is addressed in Chapter Three, and further expanded upon in Chapter Four: the issue of the relative merit of generating MS^3 data at all for phosphopeptide analysis. Despite the fact that it is an approach chosen by many if not most of the investigators in this area, there is controversy as to the value of these spectra in providing more information for peptide identification than simply generating MS^2 spectra alone. It had been suggested that the duty cycle time spent generating MS^3 spectra on the instrument results in fewer unique identifications than an equivalent run generating MS^2 spectra only.

Chapter Four presents a full analysis of several methods for analyzing a highly phosphopeptide-enriched sample on an ion-trap instrument with a high-accuracy Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer, a Thermo LTQ-FT. The study compares results using three different methodologies: MS^2-only, the MS^2/MS^3 methodology, and a more recent Multistage Activation (MSA) methodology that is beginning to become more widely available. The project is the first critical comparison of these methodologies, assessing their relative performance using several metrics: the number of unique peptide identifications (phospho and non-phospho), the ability of the various methods to localize specific sites of phosphorylation, and whether method choice has an impact on the overall mass accuracy of data produced. Also compared are three different informatics strategies for processing the MS^2 and MS^3 spectra. Overall, the

study attempts to define an optimal instrument and computational approach for phosphopeptide analysis.

Proteomics is a rapidly-changing area of research. Newer instrumentation methodologies and types of data are continually being produced, requiring different and increasingly sophisticated analysis methodologies. Overall, the work presented here demonstrates that informatics approaches are critical for effectively utilizing high-throughput mass spectrometry data. These results suggest that current popular software tools for analyzing these data are only using a portion of all the available information contained in spectra, and that there is significant potential for improving upon the results. The fact that proteomics is beginning to have a serious impact on problems of significant clinical interest, providing very valuable data that can be used in the treatment of disease, underscores the importance of developing optimal strategies for processing and mining these data. Phosphoproteomics in particular is of immense importance in the understanding the dynamics of diseases such as cancer. Signaling pathways directing cellular growth and differentiation are almost always altered, making the elucidation of the components in this process critical. The work discussed below improves upon the ability to characterize these components.

References

1. Dayhoff,M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD.

2. Smith, T. F., Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol.*147: 195-7.

3. Kohara, Y., Akiyama, K., Isono, K. (1987) The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 50: 495-508.

4. Goffeau, A., et al. (1996) Life with 6000 genes. *Science*. 274: 546, 563-7.

5. Venter, J.C., et al. (2001) The sequence of the human genome. *Science.* 291: 1304-51.

6. Lander, E.S., et al. (2001) Initial sequencing and analysis of the human genome. *Nature.* 409: 860-921.

7. O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250: 4007-21.

8. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: a novel approach to testing for induced point mutations in mammals. *Humangenetik* 26: 231-243.

9. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C.,Grimley, C.,Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence database, PNAS 90: 5011-5015.

10. Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E., Engelhard, V. H. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255: 1261-1263.

11. Yates, J.R., III, McCormack, A.L., Schieltz, D., Carmack, E. Link, A. (1997) Direct analysis of protein mixtures by tandem mass spectrometry. *J. Prot. Chem.* **16:** 495-497.

12. Wolters, D.A., Washburn, M.P.,Yates, J.R. III. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73: 5683-5690.

13. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R. (1999) Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17, 676–682.

14. Spahr, C. S., Davis, M. T., McGinley, M. D., Robinson, J.H., Bures, E.J., Beierle, J., Mort, J., Courchesne, P.L., Chen, K., Wahl, R. C., Yu, W., Luethy, R., Patterson, S. D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest. *Proteomics* 1: 93-107.

15. Washburn, M.P., Wolters, D., Yates, J.R.III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotech*. 19: 242-247.

16. Sanders, S.L., Jennings, J., Canutescu, A., Link, A.J., Weil, P.A. (2002) Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol. Cell. Biol*. 22L: 4723-4738.

17. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol*. 19: 1720-1730.

18. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17: 994-999.

19. Ross, P. L., et al. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3: 1154-69.

20. Choe, L., D'Ascenzo, M., Relkin, N. R., Pappin, D., Ross, P., Williamson, B., Guertin, S., Pribil, P., Lee, K. H. (2007) 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* 7: 3651-60.

21. Yao, X., Freas, A., Ramirez, J., Demirev, P.A., Fenselau, C. (2001) Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem.* 73: 2836-42.

22. Ong, S. E., Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol.* 1: 252-62.

23. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics*. 4: 1487-502.

24. Liu, H., Sadygov, R. G., Yates, J. R. 3rd. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem.* 76: 4193-201.

25. Spahr, C. S., Susin, S. A., Bures, E. J., Robinson, J. H., Davis, M. T., McGinley, M. D., Kroemer, G., Patterson, S. D. (2000) Simplification of complex peptide mixtures for proteomic analysis: reversible biotinylation of cysteinyl peptides. *Electrophoresis* 21:1635-50.

26. Hayes, B.K., Greis, K.D. & Hart, G.W. (1995) Specific isolation of O-Linked N-acetylglucosamine glycopeptides from complex mixtures. *Anal. Biochem*. 228: 115-122.

27. Gavin, A.C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141-147.

28. Ho, Y. et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415: 180-183.

29. Wolf-Yadlin, A.; Hautaniemi, S.; Lauffenburger, D. A.; White, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc Natl Acad Sci U S A* 104: 5860-5.

30. Keshishian, H.; Addona, T.; Burgess, M.; Kuhn, E.; Carr, S. A.(2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol Cell Proteomics* 6: 2212-29.

31. Hunter, T. (1995) Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* 80, 225 –236.

32. Cohen, P. (2000) The regulation of protein function by multisite phosphorylation– a 25 year update. *Trends Biochem. Sci.* 25: 596–601.

33. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* 298, 1912-34.

34. Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., Manning, G. (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A.* 101, 11707-12.

35. Kersten, B., Agrawal, G. K., Iwahashi, H., Rakwal, R. (2006) Plant phosphoproteomics: a long road ahead. *Proteomics* 6: 5517-28.

36. Mackay, H. J., Twelves, C. J. (2007) Targeting the protein kinase C family: are we there yet? *Nat Rev Cancer* 7: 554-62.

37. Zolnierowicz, S., and Bollen, M. (2000) Protein phosphorylation and protein phosphatases. *EMBO J.* 19: 483 –488.

38. Rikova, K., et al. (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* 131: 1190-203.

39. Guo, A., et al. (2008) Signaling networks assembled by oncogenic EGFR and c-Met. *Proc Natl Acad Sci U S A.* 105: 692-7.

40. Kewalramani, G., An, D., Kim, M. S., Ghosh, S., Qi, D., Abrahani, A., Pulinilkunnil, T., Sharma, V., Wambolt, R. B., Allard, M. F., Innis, S. M., Rodrigues, B. (2007) AMPK control of myocardial fatty acid metabolism fluctuates with the intensity of insulin-deficient diabetes. *J Mol Cell Cardiol.* 42: 333-42.

41. Mazanetz, M. P., Fischer PM. (2007) Untangling tau hyperphosphorylation in drug design for neurodegenerative diseases. *Nat Rev Drug Discov*. 6: 464-79.

42. Loyet,K. M., Stults, J. T., Arnott, D. (2005) Mass spectrometric contributions to the practice of phosphorylation site mapping through 2003: a literature review. *Mol Cell Proteomics* 4: 235-45.

43. Li, X., Gerber, S. A., Rudner, A. D., Beausoleil, S. A., Haas, W., Villén, J., Elias, J. E., Gygi, S.P. (2007) Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J Proteome Res.* 6, 1190-7.

44. Chi, A., Huttenhower, C., Geer, L. Y., Coon, J. J., Syka, J. E., Bai, D. L., Shabanowitz, J., Burke, D. J., Troyanskaya, O. G., Hunt, D. F. (2007) Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry *Proc Natl Acad Sci U S A.* 104: 2193-8.

45. Molina, H., Horn, D. M., Tang, N., Mathivanan, S., Pandey, A. (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A.* 104: 2199-204.

46. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., and Aebersold, R. (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nature methods* 4, 231-237.

47. Tholey, A., Reed, J., Lehmann, W. D. (1999) Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J Mass Spectrom*. 34: 117-23.

48. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 101, 13417-13422

49. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 4, 310-327.

50. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 101, 12130-12135

51. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell.* 127: 635-48.

52. Macek, B., Mijakovic, I., Olsen, J.V., Gnad, F., Kumar, C., Jensen, P. R., Mann, M. (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium *Bacillus subtilis. Mol Cell Proteomics* 6: 697-707.

53. Hoffert, J. D., Wang, G., Pisitkun, T., Shen, R.F., Knepper, M. A. (2007) An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins. *J Proteome Res.* 6: 3501-8.

54. Ulintz, P. J., Bodenmiller, B., Andrews, P. C., Aebersold, R., Nesvizhskii, A. I. (2008) Investigating MS^2-MS^3 matching statistics: A model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. *Mol Cell Proteomics* 7, 71-87.

55. Wu, J., Shakey, Q., Liu, W., Schuller, A., Follettie, M. T. (2007) Global profiling of phosphopeptides by titania affinity enrichment. *J Proteome Res.* 6: 4684-9.

56. Palumbo, A. M., Tepe, J. J., Reid, G. E. (2008) Mechanistic Insights into the Multistage Gas-Phase Fragmentation Behavior of Phosphoserine- and Phosphothreonine-Containing Peptides. *J Proteome Res.* Jan 9; [Epub ahead of print]

57. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc *Natl Acad Sci U S A.* 101: 9528-33.

58. Johnson, R. S., Biemann, K. (1989) Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed Environ Mass Spectrom.* 18: 945-57.

59. Bartels, C. (1990) Fast algorithm for peptide sequencing by mass spectrometry. *Biomed Environ Mass Spectrom.* 19: 363-68.

60. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 90: 5011-5015.

61. Mann, M., Hojrup, P., Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338-345.

62. Pappin, D.J.C., Hojrup, P. & Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3: 327-332.

63. James, P., Quadroni, M., Carafoli, E, Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 195: 58-64.

64. Yates, J.R., III, Speicher, S., Griffin, P. R., Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 214, 397-408.

65. Eng, J.K., McCormack, A.L.,Yates, J.R., III. (1994) An approach to correlate tandem mass spectral data pf peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5: 976-989.

66. Mann, M., Wilm, N. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 66: 4390-4399.

67. Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 20: 3551-67.

68. Clauser, K. R., Baker, P., Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*. 71: 2871-82.

69. Zhang, W., Chait, B. T. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem.* 72: 2482-9.

70. Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 112: 531-52.

71. Taylor, J. A., Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 11: 1067-75.

72. Dancík, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A. (1999) De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol.* 6:327-42.

73. Ma, B., Zhang ,K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 17: 2337-42.

74. Frank, A., Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem.* 77: 964-73.

75. Hernandez, P., Gras, R., Frey, J., Appel, R.D. (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* 3: 870–878.

76. Bandeira, N., Clauser, K. R., Pevzner, P. A. (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol Cell Proteomics*. 6: 1123-34.

77. Zimmer JS, Monroe ME, Qian WJ, Smith RD. (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev.* 25: 450-82.

78. Lipton, M. S., et al (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci U S A.* 99: 11049-54.

79. Nesvizhskii, A. I., Vitek, O., Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*. 4: 787-97.

80. Eriksson, J., Chait, B. T., Fenyö, D. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal Chem*. 72: 999-1005.

81. Fenyö, D., Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem.* 75: 768-74.

82. Sadygov, R. G., Yates, J. R. 3rd. (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem.* 75: 3792-8.

83. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi ,W., Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J Proteome Res.* 3: 958-64.

84. Keller, A., Nesvizhskii, A. I., Kolker, E. Aebersold, R.. (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem*. 74: 5383-5392.

85. Choi H, Nesvizhskii AI. Semisupervised model-based validation of Peptide identifications in mass spectrometry-based proteomics. J Proteome Res. 2008 Jan;7(1):254-65.

86. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2, 43-50.

87. Elias, J. E., Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. Mar 4: 207-14.

88. Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem*. **75**, 4646-58.

89. Craig, R., Cortens, J.C., Fenyo, D. Beavis, R. C. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res*. 5: 1843–1849.

90. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S.,MacCoss, M. J. (2007) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem*. 78: 5678–5684.

91. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., Aebersold, R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7: 655–667.

92. Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., Pevzner, P. A. (2008) Clustering millions of tandem mass spectra. *J Proteome Res*. 7: 113-22.

93. Bandeira, N., Tang, H., Bafna, V., Pevzner, P. Shotgun protein sequencing by tandem mass spectra assembly. *Anal Chem*. 76: 7221-33.

94. Flikka, K., Meukens, J., Helsens, K., Vandekerckhove, J., Eidhammer, I., Gevaert, K., Martens, L. (2007) Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *Proteomics* 7: 3245-58.

95. Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., VerBerkmoes, N. C., McDonald, W. H. (2005) MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J Am Soc Mass Spectrom*. 16: 1250-61.

96. Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*. 23: 1562-7.

97. Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A., Tang, W. H., Keating, S. P., Hunter, C. L., Nuwaysir, L. M., Schaeffer, D. A. (2007) The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature

probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics* 6:1638-55.

98. Siuti, N., Kelleher, N. L. (2007) Decoding protein modifications using top-down mass spectrometry. *Nat Methods.* 4: 817-21

99. Bogdanov, B., Smith, R. D. (2005) Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom Rev*. 24: 168-200.

100. Falkner JA, Falkner JW, Andrews PC. ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats. Bioinformatics. 2007 Jan 15;23(2):262-3.

101. Falkner JA, Falkner JW, Andrews PC. ProteomeCommons.org JAF: reference information and tools for proteomics. Bioinformatics. 2006 Mar 1;22(5):632-3.

102. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 77, 4626-4639.

Chapter 2
Improved classification of mass spectrometry database search results using machine learning approaches[5]

2.1 Summary

Manual analysis of mass spectrometry data is a current bottleneck in high-throughput proteomics. In particular, the need to manually validate the results of mass spectrometry database searching algorithms can be prohibitively time-consuming. Development of software tools that attempt to quantify the confidence in the assignment of a protein or peptide identity to a mass spectrum is an area of active interest. The goal of this study is to extend work in this area by investigating the potential of recent machine learning algorithms to improve the accuracy of these approaches, and as a flexible framework for accommodating new data features. Specifically, the ability of boosting and random forest approaches to improve the discrimination of true hits from false positive identifications in the results of mass spectrometry database search engines, compared to thresholding and other machine learning approaches, is demonstrated. We accommodate additional attributes obtainable from database search results, including a

[5] The study outlined in the chapter was originally published by Ulintz PJ, Zhu J, Qin ZS, Andrews PC as "Improved classification of mass spectrometry database search results using newer machine learning approaches." Mol Cell Proteomics. 2006 Mar;5(3):497-509.

factor addressing proton mobility. Performance is evaluated using publically available electrospray data and a new collection of MALDI data generated from purified human reference proteins.

2.2 Introduction

The proteomics field is driven by the need to develop increasingly high-throughput methods for the identification and characterization of proteins. Mass spectrometry (MS) is the primary experimental method for protein identification; tandem mass spectrometry (MS/MS) in particular is now the de-facto standard identification technology, providing the ability to rapidly characterize thousands of peptides in a complex mixture. Instrument development continues to improve the sensitivity, accuracy and throughput of analysis. Current instruments are capable of routinely generating several thousand spectra per day, detecting sub-femtomolar levels of peptide at 10 ppm mass accuracy or better. Such an increase in instrument performance is limited, however, without effective tools for automated data analysis. In fact, the primary bottleneck in high-throughput proteomics production 'pipelines' is in many cases no longer the rate at which the instrument can generate data, but rather it is in quality analysis and interpretation of the results to generate confident protein assignments. This bottleneck is primarily due to the fact that it is often difficult to distinguish true hits from false positives in the results generated by automated mass spectrometry database search algorithms. All MS database search approaches produce scores describing how well a peptide sequence matches experimental fragmentation, yet classifying hits as “correct” or “incorrect” based on a simple score threshold frequently produces unacceptable false positive/false negative rates. Consequently, manual validation is often required to be truly confident in the assignment of a database protein to a spectrum.

Software and heuristics for automated and accurate spectral identification (1-7) and discrimination of correct and incorrect hits (8-16) is thus an onging effort in the proteomics community, with the ultimate goal being completely automated MS data interpretation. The most straightforward approach to automated analysis is to define specific score-based filtering thresholds as discriminators of correctness, e.g. accepting

SEQUEST scores of doubly-charged fully tryptic peptides with XCorr > 2.2 and delta Cn values at least 0.1 (17); these thresholds are typically published as the criteria for which correctness is defined. Other efforts have focused on establishing statistical methods for inferring the likelihood that a given hit is a random event. A well known example of this is the significance threshold calculated by the Mascot search algorithm, which by default displays a threshold indicating the predicted probability of an assignment being greater than 5% likely to be a false positive based on the size of the database. Use of a reverse database search to provide a measure of false positive rate is another method frequently used (8, 18). More formally, Sadygov and Yates model the frequency of fragment ion matches from a peptide sequence database matching a spectrum as a hypergeometric distribution (12), a model also incorporated into the openly available X!Tandem algorithm (6, 13); while Geer et al. model this distribution as a Poisson distribution (7).

Keller, et al. (11) were among the first to implement a generic tool for classifying the results of common search algorithms as either correct or incorrect. Their PeptideProphet tool represents arguably the most well-known openly-available tool implementing a probabilistic approach to assess the validity of peptide assignments generated by MS database search algorithms. Their approach contains elements of both supervised and unsupervised learning, achieving a much higher sensitivity than conventional methods based on scoring thresholds. One concern with PeptideProphet, however, is the degree to which the supervised component of the model can be generalized to new types of data and the ease with which new potentially useful information can be added to the algorithm.

This work attempts to address these difficulties by applying a set of simple "over the counter" methods to the challenging peptide identification problem. Anderson, et al. demonstrated that support vector machines could perform well on ion-trap spectra searched using the SEQUEST algorithm (14). In this chapter, we demonstrate that the latest machine learning techniques for classification, namely, tree-based ensemble methods such as boosting and random forest, are more suitable for the peptide classification problem and provide improved classification accuracy. The rationale for the improvements lies in their ability to efficiently combine information from multiple easy-to-get, but dependent and weakly discriminatory attributes. Such work will hopefully

result in development of software tools that are easily installed in a production laboratory setting that would allow convenient filtering of false identifications with an acceptably high accuracy, either as new tools or as a complement to currently existing software. The problem of classification of mass spectrometry-based peptide identification seems well suited to these algorithms and could lead to more readily-usable software for automated analysis of the results of mass spectrometry experiments.

2.3 Experimental Procedures

2.3.1 Overview of classification techniques

Mixture Model Approach in PeptideProphet

Among all the methods that have been proposed in the literature for the peptide identification problem, the mixture model approach implemented in the PeptideProphet algorithm (11) is perhaps the most well known. In this method, a discriminant score function $F(x_1, x_2, ..., x_s) = c_0 + c_1 x_1 + ... + c_s x_s$ is defined to combine database search scores x_1, x_2,...,x_S where c_i's are weights. Based on a training dataset, a Gaussian distribution is chosen to model the discriminant scores corresponding to correct peptide assignments, and a Gamma distribution is selected to model the asymmetric discriminant scores corresponding to incorrect peptide assignments. All the scores are therefore represented by a mixture model $p(x) = rf_1(x) + (1-r)f_2(x)$, where $f_1(x)$ and $f_2(x)$ represent the density functions of the two types of discriminant scores, and r is the proportion of correct peptide identifications. For each new test dataset, the EM algorithm (19) is used to estimate the probability that the peptide identified is correct. A decision can be made by comparing the probability to a pre-specified threshold. When compared to conventional means of filtering data based on SEQUEST scores and other criteria, the mixture model approach achieves much higher sensitivity.

A crucial part of the above approach is the choice of discriminant score function F. In (11), the c_i's are derived in order to maximize the between- versus within-class variation under the multivariate normal assumption using training data. To make this method work, one has to assume that the training data and the test data are generated from

the same source. When a new set of discriminant scores is generated and needs to be classified, one has to retrain the c_i weight parameters using a new corresponding training set; in other words, the discriminant function F is data dependent. In an area such as proteomics in which there is a good amount of heterogeneity in instrumentation, protocol, database, and database searching software, it is fairly common to come across data which display significant differences. It is unclear to what degree the results of a classification algorithm are sensitive to these differences, hence it is desirable to automate the discriminant function training step. Another potential issue is the Normal and Gamma distribution used to model the two types of discriminant scores. There is no theoretical explanation why the discriminant scores should follow these two distributions; in fact, a Gamma distribution rather than a Normal distribution may be appropriate for both positive and negative scores when using the Mascot algorithm (20). It is possible that for a new set of data generated by different mass spectrometers and/or different search algorithms, the two distributions may not fit the discriminant scores well. Also, certain types of data attributes or scores may be more difficult to accommodate into such a model if those attributes significantly alter the shape of the discriminant score distribution. For example, qualitative or discrete attributes may be more difficult to model. As a result, higher classification errors may be produced using this model-based approach.

Machine learning techniques

Distinguishing correct from incorrect peptide assignments can be regarded as a classification problem, or supervised learning, a major topic in the statistical learning field. Many powerful methods have been developed such as CART, SVM, random forest, boosting, and bagging (21). Each of these approaches has unique features that enable them to perform well in certain scenarios; the SVM, for example, is a good tool for small sample size, large feature space situations. On the other hand, all approaches are quite flexible and have been applied to an array of biomedical problems. In this project, several of these state-of-the-art machine learning approaches are applied to the peptide assignment problem.

Boosting

The boosting idea, first introduced by Freund and Schapire with their AdaBoost algorithm (22), is one of the most powerful learning techniques introduced during the past decade. It is a procedure that combines many "weak" classifiers to achieve a final powerful classifier. This section provides a concise description of boosting in the two-class classification setting. Suppose we have a set of training samples, where x_i is a vector of input variables—in this case, various scores and attributes of an individual MS database search result produced from an algorithm such as SEQUEST— and y_i is the output variable coded as -1 or 1, indicating whether the sample is an incorrect or correct assignment of a database peptide to a spectrum. Assume we have an algorithm that can build a classifier $T(x)$ using weighted training samples so that, when given a new input x, $T(x)$ produces a prediction taking one of the two values {-1, 1}; the classifier $T(x)$ is typically a decision tree. Then boosting proceeds as follows: start with equal weighted training samples and build a classifier $T_1(x)$. If a training sample is misclassified, e.g. an incorrect peptide is assigned to the spectrum, the weight of that sample is increased (boosted). A second classifier $T_2(x)$ is then built with the training samples, but using the new weights, no longer equal. Again, misclassified samples have their weights boosted and the procedure is repeated M times. Typically, one may build hundreds or thousands of classifiers this way. A final score is then assigned to any input x, defined to be a linear (weighted) combination of the classifiers. A high score indicates that the sample is most likely a correctly assigned protein with a low score indicating that it is most likely an incorrect hit. By choosing a particular value of the score as a threshold, one can select a desired specificity or a desired ratio of correct to incorrect assignments.

Random Forests

Similar to boosting, the random forest (23) is also an ensemble method that combines many decision trees. However, there are three primary differences in how the trees are grown: 1. Instead of assigning different weights to the training samples, the method randomly selects, with replacement, n samples from the original training data; 2. Instead of considering all input variables at each split of the decision tree, a small group of input variables on which to split are randomly selected; 3. Each tree is grown to the largest extent possible. To classify a new sample from an input, one runs the input down

each of the trees in the forest. Each tree gives a classification (vote). The forest chooses the classification having the most votes over all the trees in the forest. The random forest enjoys several nice features: like boosting, it is robust with respect to input variable noise and overfitting, and it gives estimates of what variables are important in the classification. A discussion of the relative importance of the different attributes used in our analysis of MS search results is given in the results section.

Support vector machines

The support vector machine (SVM) is another successful learning technique (24). It typically produces a non-linear classification boundary in the original input space by constructing a linear boundary in a transformed version of the original input space. The dimension of the transformed space can be very large, even infinite in some cases. This seemingly prohibitive computation is achieved through a positive definite reproducing kernel, which gives the inner product in the transformed space. The SVM also has a nice geometrical interpretation in the finding of a hyperplane in the transformed space that separates two classes by the biggest margin in the training samples, although this is usually only an approximate statement due to a cost parameter. The SVM has been successfully applied to diverse scientific and engineering problems, including the life sciences (25, 26, 27). Anderson, et al. (14) introduced the SVM to MS/MS spectra analysis, classifying SEQUEST results as correct and incorrect peptide assignments. Their result indicates that the SVM yields less false positives and false negatives compared to other cutoff approaches.

However, one weakness of the SVM is that it only estimates the category of the classification, while the assignment probability $p(x)$ may be of interest itself, where $p(x) = P(Y = 1 \mid X = x)$ is the posterior probability of a sample being in class 1 (i.e. a correctly identified peptide). Another problem with the SVM is that it is not trivial to select the best tuning parameters for the kernel and the cost. Often a grid search scheme has to be employed, which can be time consuming. In comparison, boosting and the random forest are very robust, and the amount of tuning needed is rather modest compared with the SVM.

2.3.2 Reference Datasets

Two collections of mass spectrometry data were used in this study, representing the two most common protein MS ionization approaches: electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). The performance of boosting and random forest methods in comparison with other approaches using a known, published ESI dataset is benchmarked. The performance these methods to newer in-house generated MALDI data from an ABI-TOF/TOF (28) are then evaluated.

ESI-SEQUEST dataset

The electrospray dataset was kindly provided by Andy Keller, as described in (11) and (29). These data are combined MS/MS spectra generated from twenty-two different LC-MS/MS runs on a control sample of eighteen known (non-human) proteins mixed in varying concentrations. A ThermoFinnigan ion trap mass spectrometer was used to generate the dataset. In total, the data consists of 37,044 spectra of three parent ion charge states: $[M+H]^{+}$, $[M+2H]^{2+}$ and $[M+3H]^{3+}$. Each spectrum was searched by SEQUEST against a human protein database with the known protein sequences appended. The top-scoring peptide hit was retained for each spectrum; top hits against the known eighteen proteins were labeled as "correct", and manually verified by Keller, et al. All peptide assignments corresponding to proteins other than the eighteen in the standard sample mixture and common contaminants were labeled as "incorrect". In all, 2757 (7.44%) peptide assignments were determined to be correct. The distribution of hits as used to train and test the methods used in this study are indicated in Table 1.

MALDI-SpectrumMill dataset

One goal of the study was to evaluate performance of the algorithms on data generated using different instrumentation; namely, MALDI data. Toward that end, 300 purified recombinant human protein samples were procured from Genway Biotech Inc (San Diego, CA). Aliquots of these proteins were run on 1D SDS-PAGE to confirm

purity and protein molecular weight. Plugs from the band on each 1D gel were subjected to in-gel trypsin digestion, serially diluted (in many cases) to generate potentially more 'realistic' spectra, and cleaned by C-18 ZipTip. Resulting digestions for each protein were spotted in four replicates on MALDI target plates, and MS/MS spectra acquired on an Applied Biosystems Inc (Foster City, CA) 4700 Proteomics Analyzer ("TOF/TOF"). Spectra were collected from the successive replicate spots by selecting the most abundant ions from each replicate and excluding previously selected peaks, until reasonably sized MS peaks could no longer be detected; this process resulted in up to twenty-four MS/MS spectra for each standard protein. At the time of this analysis, results from 158 of these protein standards had been generated and were used to compose the dataset.

To generate testing and training datasets for this analysis, all MALDI spectra were searched using the Agilent (Palo Alto, CA) Spectrum Mill platform (specifying trypsin as the proteolytic enzyme, and accommodating oxidized methionine and Pyro-Glu modifications) against a version of the NCBInr-Human dataset downloaded March 10, 2005. The NCBInr database was modified by replacing all entries corresponding to the Genway protein standards with annotated entries that include the appropriate N-term affinity tag (either T7 or 6xHis), and appending the *E. coli* K12 sequences, as *E. coli* was the host organism in which the recombinants were produced. Tolerances of 0.7 Da for precursor ion selection and 0.3 Da for fragment ion selection were used in the search, with a minimum matched peak intensity of 40%. "Correct/Incorrect" labels were assigned to each spectrum in a semi-automated manner by correlating the accession number of the search result with that of the protein digest known to be spotted at the plate location from which each spectrum was derived. The dataset consists of 11764 search results from 4340 spectra (up to the top five ranking hits for each spectrum). 1044 are 'true positive' hits; of these, 111 are non-top-ranking hits. The precise distribution of hits as used to test the algorithms are shown in Table 1. Note, the relatively low fraction of true positives as per what would be expected from this instrument is primarily due to the fact that the top five ranking hits were selected, not just the top hits, from every search result. The inclusion of non-top ranking hits was done to examine whether the machine learning tools could be used to distinguish correct hits among these lower-ranked results,

a frequent occurrence in MS/MS search data. The lower ranked hits contain potentially valuable protein identifications that would be discarded using many normal approaches.

The complete set of annotated MALDI protein standard data (named 'Aurum') is available for download on the ProteomeCommons website[6]. To our knowledge, this represents to first publicly-available annotated MALDI dataset useful for development of these types of algorithms.

	Training	Testing
ESI-SEQUEST		
Correct	1930	827
Incorrect	24001	10286
Total	25931	11113
MALDI-SpectrumMill		
Correct	731	313
Incorrect	7504	3216
Total	8235	3529

Table 2-1. Number of spectra examples for the ESI-SEQUEST dataset and the MALDI-Spectrum Mill dataset

2.3.3 Attributes extracted from each dataset

Both *.out search result files from SEQUEST and *.spo result files from Spectrum Mill were parsed into a simple text row/column format suitable for use by pattern classification algorithms using custom modules written in Python (available upon request). For the SEQUEST results, only the top hit for each spectrum was parsed; again, for the MALDI dataset, the top five ranking hits were retained.

SEQUEST

[6] http://www.proteomecommons.org

The attributes extracted from SEQUEST assignments are listed in Table 2. Attributes include typical scores generated by the SEQUEST algorithm (Sp, Sp rank, deltaCn, XCorr), as well as other statistics included in a SEQUEST report (total intensity, number of matching peaks, fragment ions ratio). Length is included amongst the PeptideProphet attributes since PeptideProphet normalizes the XCorr attribute using the length of the peptide. Number of tryptic termini (NTT) is a useful measure for search results obtained by specifying no proteolytic enzyme, and is used extensively in PeptideProphet (11). Other attributes include features readily obtainable from the candidate peptide sequence: C-term residue (K='1', R='2', others='0'), number of prolines, and number of arginines. A new statistic, the Mobile Proton Factor (MPF), is calculated as:

$$\frac{(1.0 * R) + (0.8 * K) + (0.5 * H)}{Ch\arg e}$$

MPF attempts to provide a simple measure of the mobility of protons in a peptide, a theoretical measure of the ease of which a peptide may be fragmented in the gas phase (30, 31, 32). A smaller value for MPF is indicative of higher protein mobility, whereas peptides with MPF >= 1 can be considered 'nonmobile'. *R*, *K*, and *H* refer to the number of these amino acid residues present in the sequence, reflecting the overall basicity of the peptide. The coefficients for these factors reflect the relative basicity of the three residues normalized to the dissociation constant of arginine (the pKa values of Arg, Lys and His being 12.0, 10.0, and 5.9, respectively). Charge indicates the charge on the parent peptide, reflecting the number of free protons potentially available for charge-directed fragmentation. We include MPF to demonstrate the ease of accommodation of additional information into the classification algorithms, amounting to simply adding an additional data column to the data set.

Spectrum Mill:

Spectrum Mill attributes are indicated in Table 3. Spectrum Mill is a search platform based initially on the Protein Prospector set of scripts (3), further developed by Karl Clauser and commercialized by Agilent. The primary Spectrum Mill score is non-probabilistic, intended as an absolute measure of the information contained in an

assignment of a spectrum with a peptide sequence, and is database-size independent. The score increases with peaks matching theoretical fragment ions types (in an instrument-dependent way), with a penalty for unmatched peaks. Intensity of peaks is a factor in the score, as is peptide length. The scoring system accommodates all major fragmentation ion types and neutral losses, and the presence of internal and immonium ions. Scored Peak Intensity (SPI) is the second primary Spectrum Mill scoring measure, reflecting the percentage of total peak intensity matching predicted fragment ion masses. Spectrum Mill implements semi-automated spectrum validation and curation tools based on linear thresholds for primary Score and SPI. Background Cleavage Score (BCS) indicates the number of cleavage events generating b- or y-ions; unused ion ratio (number of unused ions / number of total ions after peak thresholding) provides a measure of the amount of signal not accounted for in the match; and delta parent mass measures the difference between the observed and experimental peptide precursor masses. Terms such as parent charge were avoided in the Spectrum Mill data due to the fact that the ions are produced by MALDI.

Attribute Group	Attribute Name	SEQUEST Name	Description
PeptideProphet **(I)**	Delta MH+	(M+H)+	Parent ion mass error between observed and theoretical
	Sp Rank	Rank/Sp	Initial peptide rank based on preliminary score
	Delta Cn	deltCn	1 – Cn: difference in normalized correlation scores between next-best and best hits
	XCorr	XCorr	Cross-correlation score between experimental and
	Length	Inferred from Peptide	Length of the peptide sequence
NTT **(II)**	Number of Tryptic Termini (NTT)	Inferred from Peptide	Measures whether the peptide is fully tryptic, partially tryptic, or non-tryptic (2, 1, or 0, respectively)
Additional **(III)**	**Parent Charge**	(+1), (+2), (+3)	Charge of the parent ion
	Total Intensity	total inten	Normalized summed intensity of peaks
	DB peptides within mass	# matched peptides	Number of database peptides matching the parent peak mass within the specified mass tolerance
	Sp	Sp	Preliminary score for a peptide match
	Ion Ratio	Ions	Fraction of theoretical peaks matched in the preliminary
	C-term Residue	Inferred from Peptide	Amino acid residue at the C-term of the peptide (1 = R, 2 = 'K', 0 = 'other')
	Number of Prolines	Inferred from Peptide	Number of prolines in the peptide
	Number of Arginines	Inferred from Peptide	Number of arginines in the peptide
Calculated **(IV)**	Proton Mobility Factor	calculated	A measure of the ratio of basic amino acids to free protons for a peptide (described in Experimental Procedures)

Table 2-2. SEQUEST attribute descriptions. Attribute names in bold are treated as discrete categorical variables. Abbreviation DB = database.

Attribute Name	Attribute Description
Delta MH+	Parent ion mass error between observed and theoretical
Rank	Rank by Score and number of unmatched ions of the peptide among all peptides matched for the spectrum
Score	SpectrumMill score
Percent Scored Peak Intensity (SPI)	Percentage of total peak intensity from observed peaks which match theoretical fragment ion masses
Backbone Cleavage Score (BCS)	Number of backbone cleavage events from which a y or a b ion is observed
Unused Ions Ratio	(Number of observed peaks not matched to fragment ion masses)/ (Number of observed peaks matched to fragment ion masses)

Table 2-3. Spectrum Mill attribute descriptions.

2.3.4 Implementation Specifics

A single training and testing dataset was constructed for each of the ESI-SEQUEST and MALDI-SpectrumMill datasets by random selection. Random sampling was done separately for "correct"-labeled and "incorrect"-labeled data so that both training and testing data contain the same proportions. For all results, evaluation was done on a test set that does not overlap the training set. Two thirds of all data were used for training and one third for testing.

The PeptideProphet standalone application used in this analysis was downloaded from peptideprophet.sourceforge.net. PeptideProphet is also available as part of the Trans-Proteomics Pipeline being developed at the Seattle Proteome Center (http://tools.proteomecenter.org/TPP.php). All SEQUEST *.out result files corresponding to each test set were placed in a separate directory and processed using the out2summary.c script to generate the PeptideProphet html input file. PeptideProphet was run by executing the runPeptideProphet script, using default parameters. PeptideProphet was not run on the MALDI-SpectrumMill dataset.

Contributed public packages for the R programming language were used for the boosting and random forest approaches, specifically the AdaBoost algorithm (22) implemented by Greg Ridgway[7], and the randomForest v4.5-8 package, an R port by Andy Liaw and Matthew Wiener of the original Fortran algorithm developed by Leo Brieman and Adele Cutler. In general, the parameters (i.e. tree size, number of trees etc.) of the random forest and boosting implementations were not fine tuned, for two reasons: classification performances of both the random forest and boosting are fairly robust to these parameters, and also because our ultimate goal is to provide a software tool that can be easily used in a production laboratory setting without a significant tuning requirement. For the AdaBoost analysis, a decision trees with forty leaves was used for the 'weak' classifier, and the number of boosting iterations (*M)* was fixed to 1000. For random forests, the default number of attributes for each tree-- one third of the total number of attributes-- was used, except for the five-variable case in which the number of attributes was fixed at two. The default number of trees in the forest is 500, and each tree in the forest was grown until the leaf is either pure or has only five samples.

For the support vector machine, a radial kernel was chosen to classify the samples, as implemented in the libSVM package (version 2.7)[8]. The radial kernel is flexible and performed well in preliminary studies. In order to select the optimal set of tuning parameters for radial kernel, a grid search scheme was adopted using a modified version of the **grid.py** python script distributed with the libSVM package. Optimal parameters are sensitive to specific training sets: for the precise results presented in this chapter, the optimal parameters were cost = 32768.0 and gamma = 3.052e-05.

[7]http://rweb.stat.umn.edu/R/library/gbm/html/gbm.html
[8]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

2.4 Results and Discussion

2.4.1 Classification Performance

Each of the machine learning approaches used produces an ordering of the collection of examples in the test dataset. With PeptideProphet, the examples are ordered highest to lowest on the basis of a Bayesian posterior probability, as described in (11). For either boosting or random forest, the algorithm returns, in addition to a 'correct'/'incorrect' classification, an additional 'fitness' term. In the case of the random forest, the fitness term can be interpreted as a probability of the identification being correct. A probability score can be generated from the boosting fitness measure as well using a simple transformation. The SVM returns a classification and a measure of the distance to a distinguishing hyperplane in attribute space that can be considered a confidence measure. When examples are ordered in this way, results can be represented as a Receiver Operating Characteristic (ROC) plot, which provides a way of displaying the ratio of true positive classifications (sensitivity) to the fraction of false positives (1 – specificity) as a function of a variable test threshold. The threshold, chosen on the ranked ordering of results produced by the classifier, represents a trade-off between being able to select the true positives without selecting too many false positives. If the scoring threshold is set very high, the number of false positives can be minimized or eliminated but at the expense of missing a number of true positives; conversely, as the scoring threshold is lowered, more true positives are selected but more false positives will be included as well. The slope at any point in the ROC plot is a measure of the degree to which one group is included at the expense of the other.

The ESI-SEQUEST dataset allows the comparison of all four classification approaches: boosting, random forests, PeptideProphet, and the SVM. ROC plots showing the results of classifying correct vs incorrect peptide assignments of the ESI-SEQUEST dataset using these methods are shown in Figure 2-1A. All methods perform well on the data. As can be seen, the boosting and random forest methods provide a slight performance improvement over PeptideProphet and the SVM classification using the same six attributes. At a false positive rate of roughly 0.05%, the boosting and random forest achieves a sensitivity of 99% while PeptideProphet and SVM provide a 97-98%

sensitivity. It should be noted that, although a systematic difference of 1-2% can be seen in these results, this corresponds to a relatively small number of total spectra. Also indicated in Figure 2-1 are points corresponding to well-known attribute thresholds from several literature citations. Each point shows the sensitivity and specificity that would be obtained on the test dataset by applying these published thresholds to the SEQUEST attributes Charge, Xcorr, Delta Cn, and NTT.

Of interest is the fact that the boosting, random forest and SVM results asymptotically approach 1.0 sensitivity, whereas PeptideProphet approaches a sensitivity of about 0.98 for most of the length of the ROC curve (the PeptideProphet results do achieve 1.0 sensitivity at the very end). These results point to a set of spectra that the tools learn to discriminate differently; in the case of the test set described here, the discrepancy corresponds to fourteen spectra out of the total 11113. Eleven of these spectra are results annotated as 'correct' hits that PeptideProphet assigns as 'incorrect', and three 'incorrect' results that PeptideProphet assigns as 'correct'. The eleven spectra of the former category all represent singly charged spectra with very small SEQUEST DeltaCn values; PeptideProphet appears to have some difficulty with instances of +1 spectra in which the second-highest hit has a score very close to the top hit. The other three spectra represent an interesting case. Since the LCQ instrument on which the spectra were generated lacks the resolution to discriminate between doubly- and triply-charged parent ions, typical peaklist extraction protocols produce two identical peaklists for non-singly charged precursor masses, one each for the doubly- and triply-charged case. The database search for only one of these two peaklists should produce a correct result under normal circumstances. The cases that PeptideProphet “misses” here are an exception to this rule. They are cases in which two peptides containing identical sequence from the same protein-- a larger peptide with a +3 charge and one or more missed tryptic cleavage sites, and a smaller peptide (a subset of the first) with a +2 charge-- have the same apparent mass. For example, in one instance a parent precursor with a m/z of 1109 Da is selected, corresponding to a peptide from the CAH2_BOVINE protein. The CAH2_BOVINE peptide SSQQMLKFRTLNFNAEGEPELLMLANWR has a mass of 3324.82 Da. This peptide has two missed trypsin cleavage sites, a Lys at position seven and an Arg at

position nine. The peptide resulting from cleaving after position nine is TLNFNAEGEPELLMLANWR, with a mass of 2220 Da. The longer peptide was hit by SEQUEST for the +3 peaklist, and the shorter one for the +2 peaklist: 3326/3 ~= 2219/2 ~= 1109 Da. Following the rule that only one of the two identical peaklists generated from the 1109 Da precursor mass spectrum can be correct, the annotation provided with the ESI-SEQUEST dataset assigned the +3 spectrum as correct and the +2 spectrum as incorrect. The 'incorrect' annotation to the search result of the +2 peaklists is misleading in these cases, however. PeptideProphet distinguishes these cases, accurately providing a 'correct' classification for the +2 spectra. The other three completely supervised algorithms learn to classify these cases as incorrect based on similar examples in the training dataset.

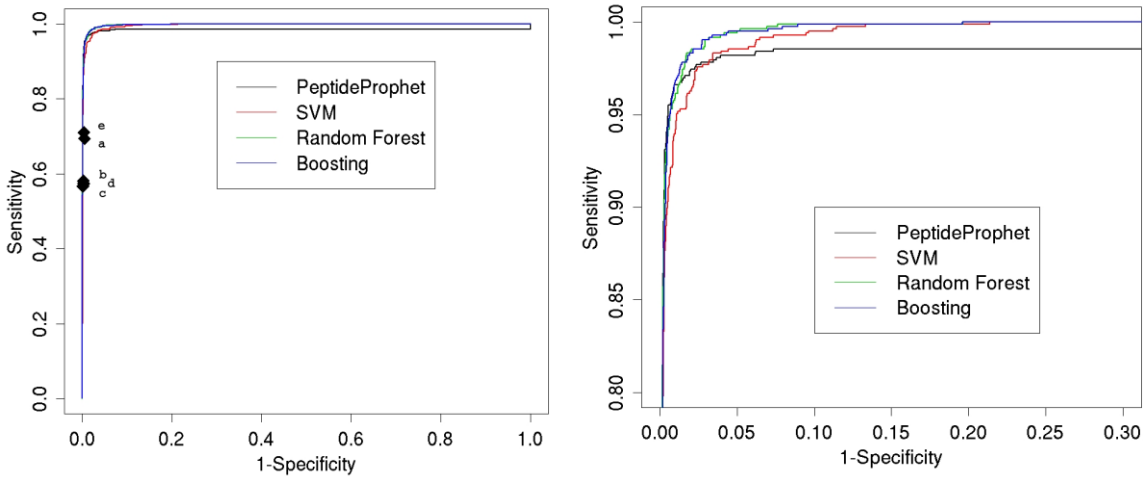
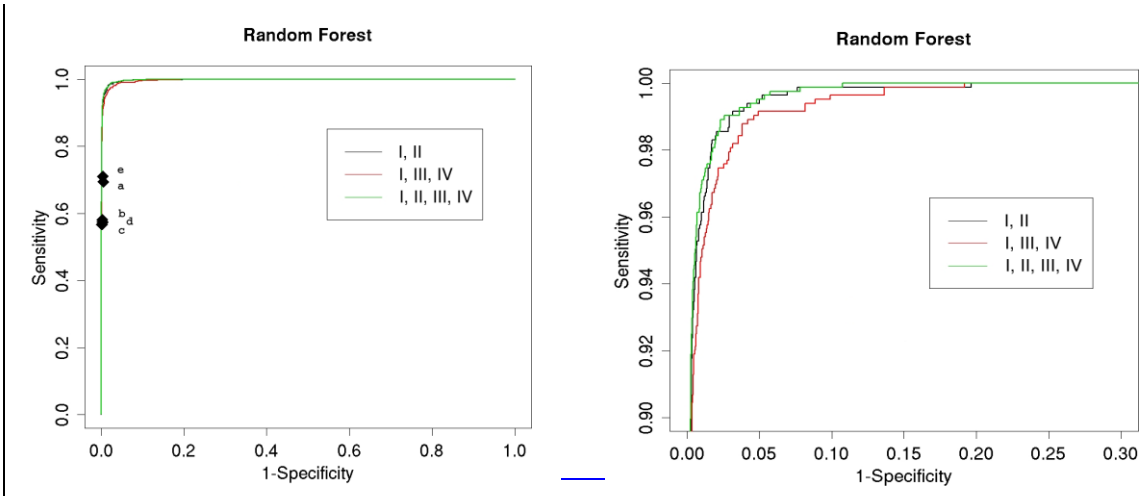
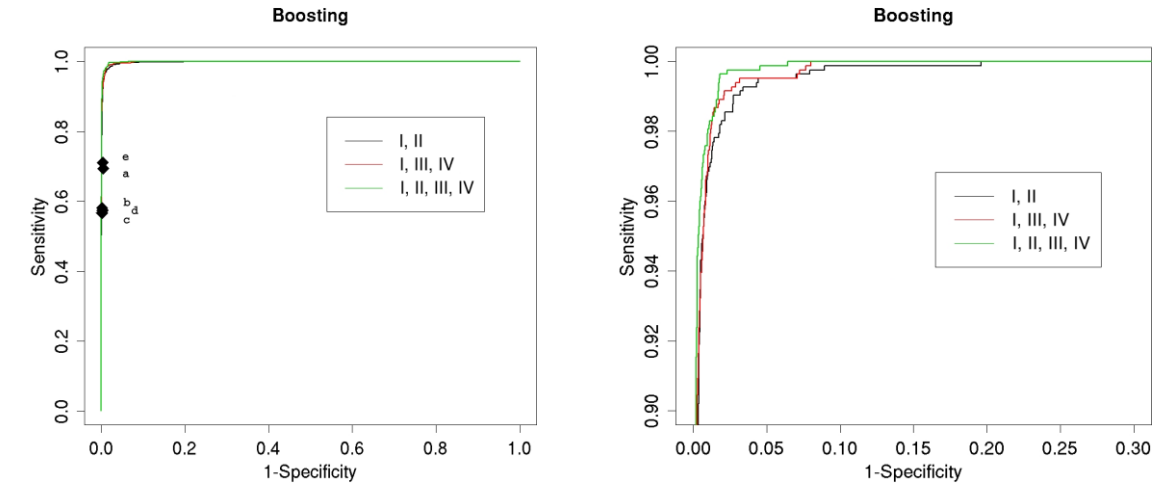


Figure 2-1. Performance machine learning methods on the ESI-SEQUEST dataset. The figure shows a ROC plot of classification of the test set by PeptideProphet, SVM, boosting, and random forest methods using attribute groups I and II. The plot on the right is a blowup of the upper left region of the figure on the left. Also displayed are points corresponding to several sets of SEQUEST scoring statistics used as linear threshold values in published studies. The following criteria were applied for choosing correct hits (the +1, +2, and +3 numbers indicate peptide charge): a, +1: XCorr ≥1.5, NTT=2; +2, +3: XCorr ≥2.0, NTT=2 (36); b, ΔCn > 0.1, +1: XCorr ≥ 1.9, NTT = 2; +2: XCorr ≥ 3 or 2.2 ≤ XCorr ≤ 3.0, NTT = 1; +3: XCorr ≥ 3.75, NTT ≥ 1 (17); c, ΔCn ≥ 0.08, +1: XCorr ≥ 1.8; +2: XCorr ≥ 2.5; +3: XCorr ≥ 3.5 (20); d, ΔCn ≥ 0.1, +1: XCorr ≥ 1.9, NTT = 2; +2: XCorr ≥ 2.2, NTT ≥ 1; +3: XCorr ≥ 3.75, NTT ≥ 1 (16); e, ΔCn ≥ 0.1, Sp rank ≤ 50, NTT ≥1, +1: not included; +2: XCorr ≥ 2.0; +3: XCorr ≥ 2.5 (11).

Panels 1B and 1C compare the performance of the boosting and random forest methods using different sets of input attributes, as shown in Table 2. The panels contains the results of these algorithms using three combinations of features: 1) attribute groups I and II: the six attributes used by the PeptideProphet algorithm (SEQUEST XCorr, Delta Cn, SpRank, Delta Parent Mass, Length, and NTT); 2) attribute groups I, III, and IV (all attributes except NTT); and 3) attribute group I-IV (all fifteen variables shown in Table 2). Overall, it can be seen that both machine learning approaches provide improvement over the scoring thresholds described in the literature. The best performance was obtained by including all fifteen variables, indicating that accommodation of additional information is beneficial. The random forest appears to be slightly more sensitive to the presence of the NTT variable than boosting. Of note is the fact that effective classification is attained by the boosting and random forest tools even in the explicit absence of the NTT variable, as demonstrated by feature combination 2), despite the fact that the ESI dataset was generated using the 'no enzyme' feature of SEQUEST. No enzyme specificity in the database search is often time-prohibitive in routine production work; it is much more common to restrict searches to tryptic peptides (or any other proteolytic enzyme used to digest the protein sample). Restricting to trypsin narrows results to having an NTT=2, rendering the attribute non-discriminatory. It must be noted, however, that in this analysis the C-term Residue attribute is not completely independent of NTT in that it contains residue information on one of the termini. If trypsin-specificity is turned on in a search, in addition to distinguishing between Lys- and Arg-terminated peptides, C-term Residue will discriminate tryptic and semi-tryptic peptides, the latter being possible if the peptide is the C-terminal peptide of a protein. If trypsin specificity is not used in the search (as in these data), although the C-term Residue variable cannot predict the NTT value of a peptide, it can discriminate between cases. If the C-term of the peptide is tryptic, the the peptide may be either fully or partially tryptic; if it is not, it can either be partially tryptic or non-tryptic.



a)



b)

Figure 2-2. Performance of Boosting and Random Forest methods using various combinations of attributes for the ESI-SEQUEST dataset. a) Results of the random forest method using various sets of attributes. The black line represents the result of the random forest using six attributes defined in Table 2-2 as groups I and II: the SEQUEST XCorr, Sp rank, ΔCn, delta parent mass, length, and NTT. The red line is the result using 14 attributes, groups I, III, and IV (no NTT). The blue line represents the result using all attribute groups I–IV, all 15 variables. **b)** ROC plot of the boosting method using attribute groups I and II (black); I, III, and IV (red); and I–IV (green). Points plotted on the curves in both full-scale figures represent published attribute threshold combinations as described in the caption for Figure 2-1.

The results of classifying the MALDI-SpectrumMill data using boosting and random forest are shown in Figure 2-2. The functionality necessary to run PeptideProphet on SpectrumMill data would require customization of the tool, and was therefore not used. The SVM on the MALDI-SpectrumMill dataset was not run since the superior performance of the boosting and random forest methods were already demonstrated on the ESI-SEQUEST dataset. Overall, the two classifiers performed similarly, with boosting outperforming random forests slightly when the false positive rate is above 15%. Both methods show dramatic improvement over thresholding combinations, based on default recommended combinations of Spectrum Mill Score and Shared Peak Intensity (SPI) values. Spectrum Mill documentation suggests three guideline threshold combinations for "Outstanding", "Good" and "Modest" hits, indicated in the figure. The "Outstanding" threshold effectively discriminates results with a low false positive rate but discards a majority of true positives. The learning approaches are able to pull a much greater number of true positives at a similar false positive rate: at a false positive rate of 5%, the learners classify roughly 70% of the true positives. The MALDI-SpectrumMill dataset was generated by selecting the top five hits from search results, since 11% of the correct results were non-top ranking hits. It does not appear to be the case that the machine learning algorithms were able to effectively select these cases out, however (data not shown). Note that the Spectrum Mill tool is intended to facilitate manual curation, and as such these threshold levels are only guidelines and are dataset-dependent not solid rules or recommended publication standards.

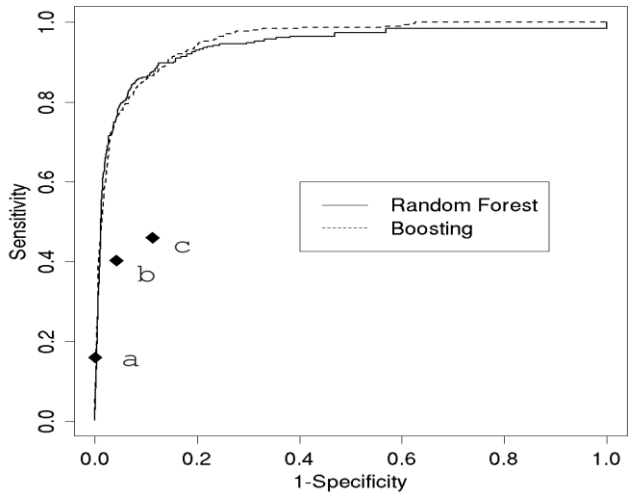


Figure 2-3. Performance of boosting and random forest methods on the MALDI-Spectrum Mill dataset. The variables used are: rank, Spectrum Mill score, SPI, delta parent mass (ΔM+H), BCS, and unused ion ratio. Points corresponding to standard guideline scoring threshold values from the Spectrum Mill documentation are displayed: **a**, Spectrum Mill score >15, SPI > 70% (outstanding); **b**, Spectrum Mill score >10, SPI > 70% (good); **c**, Spectrum Mill score >5, SPI > 70% (modest).

2.4.2 The impact of individual attributes on the final prediction accuracy

It is interesting to examine the relative importance of the various attributes used by the boosting and random forest algorithms to classify search results. The relative importance of each attribute is determined by noting nodes in the ensemble of trees in which the individual attribute appears, and summing the relative information content or loss of entropy that each node containing the attribute provides. Attributes which provide the greatest combined discrimination amongst all the nodes in which it appears thus have a higher importance.

Figure 2-4 displays the relative importance each attribute from SEQUEST and SpectrumMill search results using the boosting and random forest methods. Results for classification of the ESI-SEQUEST dataset incorporating the six attributes used by PeptideProphet are shown in Panel 2-4a. All six attributes show a contribution to the

discrimination, with the most important contribution from the NTT variable. PeptideProphet incorporates only the first five attributes in the calculation of the discriminant function, introducing NTT distributions using a separate joint probability calculation. The coefficients for their discriminant score weight XCorr highest, followed by Delta Cn, with much lower contributions due to Delta M+H and Sp Rank. Again, Length is used by PeptideProphet to correct for the well-known peptide length dependence of the XCorr variable. Our results indicate a roughly equivalent contribution from Delta Cn and XCorr, with a significant contribution from Sp Rank. Delta M+H and Length showed a much more moderate contribution. The six attributes display a high importance when used in conjunction with the other nine attributes from groups III and IV, as indicated in Panel 2-4b. Of these additional nine, Sp score shows a surprising contribution, as this scoring measure is rarely used for discrimination in popular usage. Also significant is the PeakMatchRatio measure and the MPF. For those attributes that are in common, these results are in agreement with Fisher's discriminant scores calculated by Anderson, et al. (14), with the exception of Delta M+H which showed very little contribution using their SVM approach. The number of Arginine and Proline measures, as well as Parent Charge, Length, and Number of Matched Peptides, appear to provide very little discriminative value.

The NTT variable provides by far the most important contribution, particularly for the boosting approach, but is informative only for the non-enzyme-specific searches. The results above indicate, however, that the machine learning approaches perform quite well even in the absence of this variable. The relative importances of the other measures in the absence of this variable are shown in Panel 2-4c. In this scenario, the Delta Cn measure provides the most importance contribution. A comparison of Figs 2-4b and 2-4c suggests that in the absence of NTT, the C-term Residue variable contributes much more significantly to the discrimination. As discussed above, although not as discriminatory as NTT, C-term Reside does contain some of the same information, and may be useful as a partial replacement for NTT in situations in which NTT is prohibitively time-consuming to obtain.

The relative importance of different attributes for the SpectrumMill data are shown in Figure 2-5. Not surprisingly, Rank and Score features are very important for discrimination. It is interesting, however, that Rank and Score do not duplicate each other, since the ranking of results is primarily based on Score. One surprising feature is the importance of Delta MH+ in the final discrimination. This attribute is relatively unimportant in the ESI-SEQUEST results; this may demonstrate a difference between the SpectrumMill and SEQUEST scoring schemes, but can more likely be explained by the greater mass accuracy of the TOF/TOF instrument compared to the LCQ. The percent scored peak intensity (SPI) attribute is analogous to the 'fraction matched MSMS TIC' variable published by Anderson, et al., and is one of the primary criteria used for judging the correctness of a hit in the SpectrumMill package. Its low importance measure in the context of the other attributes is of interest. The Backbone Cleavage Score and Unused Ion Ratio attributes calculated by SpectrumMill appears relatively less important in this context as well. Note, NTT was not calculated for the MALDI-SpectrumMill dataset due to difficulties in performing a no-enzyme specificity search against the NCBInr-human dataset using SpectrumMill.

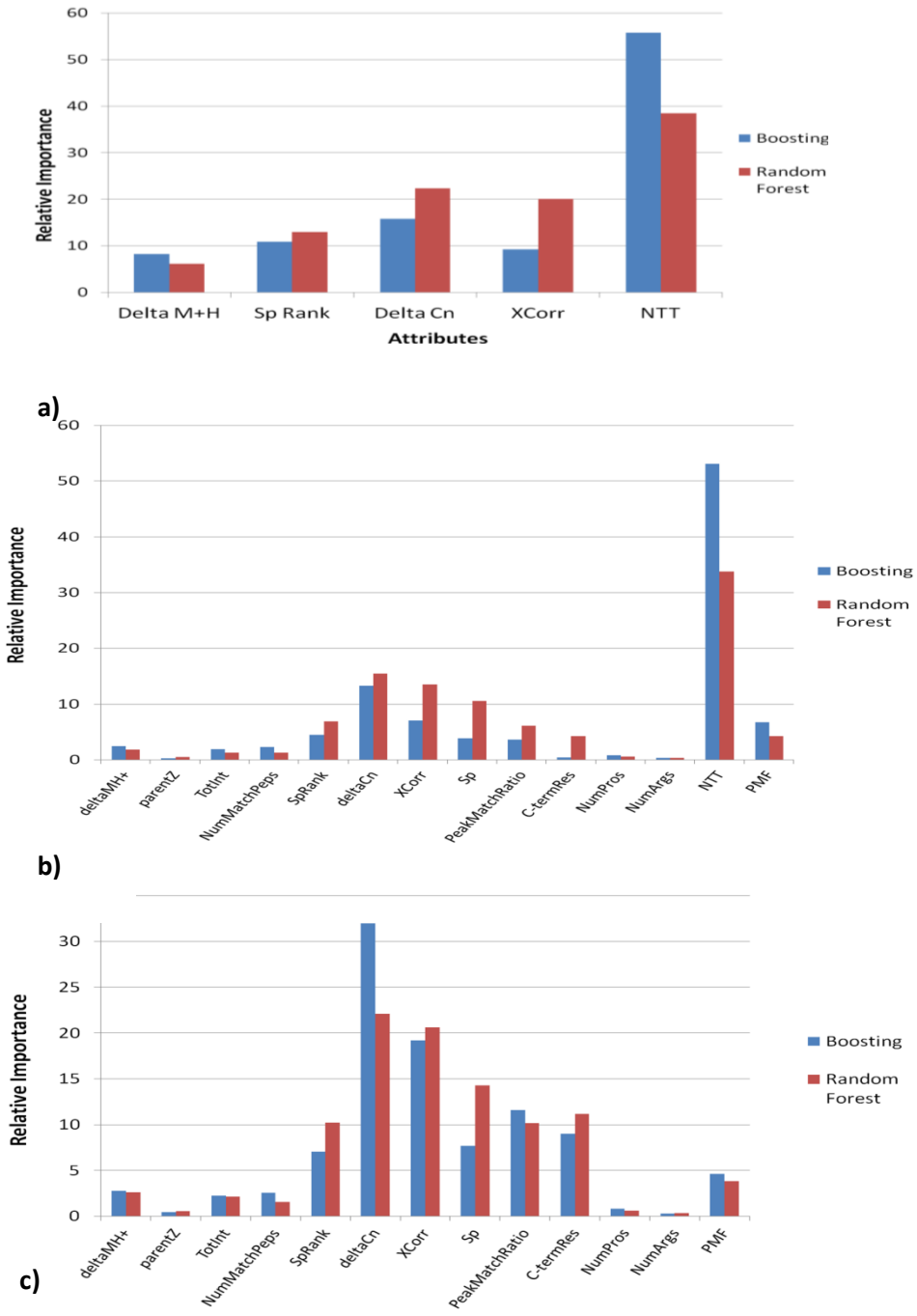


Figure. 2-4. Relative importance of SEQUEST data attributes used for classification by boosting and random forest methods. **a**) Relative importance using attribute groups I and II. **b)** Attribute importance using all attributes in random forest and boosting methods. **c)** Attribute importance using attribute groups I, III, and IV. Atribute abbreviations: TotInt, Total Intensity; NumMatchPeps, Number of Matching Peptides; NumPros, Number of Prolines; NumArgs, Number of Arginines; PMF, Protein Mobility Factor.

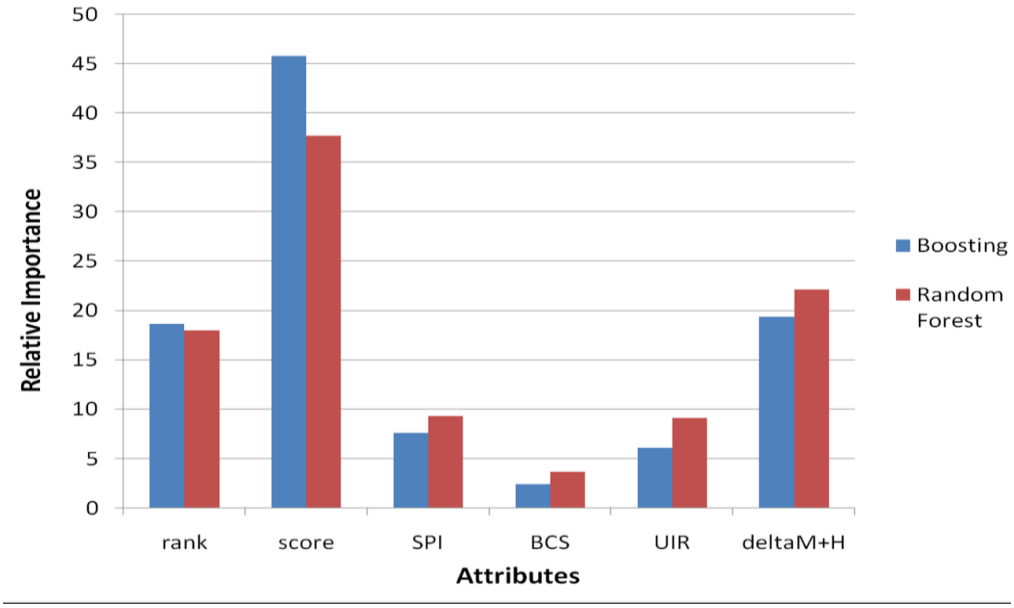


Figure 2-5. Relative importance of Spectrum Mill data attributes used for classification by boosting and random forest methods. Abbreviations: SPI, Scored Peak Intensity; BCS, Backbone Cleavage Score; UIR, Unused Ion Ratio.

2.4.3 Unsupervised Learning and Generalization/Comparison to PeptideProphet

In general, machine learning defines two primary types of models to address the classification problem, generative approaches and discriminative approaches. Algorithms such as the boosting and random forests methods discussed in this chapter are discriminative, non-parametric approaches in that they do not rely on an explicit distribution of the data, and model the posterior probability $p(y|x)$ directly (see 24 for a general description). A generative method assumes a model for the distributions of the attributes given the class label and learns the distribution $p(x|y)$, then uses Bayes rule to infer the posterior probability and the classification rule. In effect, these models incorporate assumed prior information in the form of probability distributions for each of the classes being modeled, and use these distributions to calculate the probability that a

data point belongs to each class. PeptideProphet is a generative, parametric method, modeling correct identifications as a Gaussian distribution and incorrect identifications as a Gamma distribution. If this model fits the observed data well, i.e. the distributions describing the different classes in the problem accurately reflect the physical processes by which the data are generated, the generative approach works well even for a small amount of training data. On the other hand, if the data diverge from the modeled distributions in a significant way, classification errors proportional to the degree of divergence result. Therefore, although straightforward, the performance of the parametric-based generative approaches tends to be sensitive to the assumed model. For the peptide classification problem discussed in this manuscript, there is little scientific evidence that supports a particular distribution assumption, one has to rely on past experience to make the decision. Discriminative approaches, on the other hand, are a less risky option in that they do not rely on knowledge of the distributions of classes of the data. They become increasingly safe, approaching optimality, as data size increases. Keller, et al. demonstrate that, for their data, the distributions described in their mixture model fit the data well. Whether these distributions are appropriate for all types of instruments and MS search engines, and whether they are optimal, is a research question. It may be the case that the addition of new attributes, which alter the discriminant score and thus the shape of the score distribution, may be problematic for a generative tool. Due to their flexibility, our approaches are expected to generalize well to other type of data obtained from various instruments, search engines and experimental conditions. I believe this is a particularly attractive feature.

The boosting/random forest methods are supervised approaches, relying on training data for their functionality. PeptideProphet implments both supervised and unsupervised learning components. PeptideProphet uses training data to learn coefficients in the calculation of the discriminate score; it subsequently uses these scores to establish the basic shape of the probability distributions modeling correct and incorrect search hits as a function of parent peptide charge. For each unique dataset, when additional test data arrives, the distribution parameters are refined using an EM algorithm, such that a refined classification procedure can be performed. This

unsupervised component can function to compensate for a less-than-optimal fit of observed data to the model distributions. How to combine the unsupervised learning techniques such as clustering with out-of-the-box classification tools such as the ensemble tree methods discussed here is a challenge that needs to be addressed. Our approach provides a framework for performing the supervised aspect of the problem in a more general way, using established out-of-the-box functionality. This approach can be coupled with an unsupervised component to provide more flexible functionality, assuming appropriate training datasets are available that match the input data. I have described one such dataset here, potentially useful as a distributed resource for training purposes. The degree to which an individual training dataset provides adequate parameterization for a particular test set is an open question. Certainly, training sets will need to be search algorithm specific, and it is the intention of myself and my collaborators to extend this work to other algorithms such as Mascot and X!Tandem in future studies; whether instrument-specific datasets are necessary is an area of investigation.

Having a tool which generates a truly accurate probability of correctness is an attractive feature of all algorithms in this domain. The fitness scores generated by the random forest method are probability estimates, and the boosting fitness scores can be directly converted into probability estimates via a logit transformation. True probability estimation is a difficult problem, however. It must be clearly noted that, as with other tools in this domain, these estimates must be considered approximate. Various methods have been developed for converting these types of scores into accurate probability estimates (33, 34). These methods are referred to as probability calibration methods in the literature. On the other hand, accurate classification of results does not require accurate probability estimates. For example, in a simple two-class classification setting such as ours, one may only need to know whether the probability is bigger or smaller than some selected value in order to achieve an accurate classification rule.

As a final note, the work described here addresses the problem of generating rankings and confidence measures for identification of peptides using mass spectrometry database search algorithms. This step is typically not the end-goal of data analysis: the

peptide measures can be combined to generate confidence measures for the presence of a protein in a sample. This problem of combining peptide results to generate accurate protein identifications has been addressed in other algorithms, such as ProteinProphet (35). Improvement in peptide identification will increase the sensitivity and specificity of downstream protein identifications, and the results from the algorithms described in this work can be used directly as inputs into protein-level calculations.

2.5 Conclusions

In a production proteomics lab, researchers are often faced with the challenge of curating large lists of protein identifications based on various confidence statistics generated by the search engines. The common methodology for selecting true hits from false positives is based on thresholding. These approaches can lead to a large number of false positives (using a more promiscuous threshold), or a large number of false negatives (using a relatively stringent threshold). Machine learning approaches such as boosting and random forest methods provide a more accurate method for classification of the results of MS/MS search engines as either correct or incorrect. Additionally, newer scoring criteria continue to be published which could improve the ability of automated tools to better discriminate true search results, and can complement the standard scoring measures generated by popular search engines. Flexible methods that allow for accommodation of these new scoring measures are necessary to allow them to be easily incorporated into production use. Modern machine learning approaches such as the ensemble methods described here can perform very well out-of-the box with very little tuning. Improved results could very likely be obtained by tuning these tools to particular data sets, i.e. by making use of class prior probabilities to accommodate the imbalanced sizes of the correct and incorrect datasets. These approaches can additionally be used to generate measures of relative importance of scoring variables, and may be useful in the development of new scoring approaches.

References

1. Eng, J., McCormack, A., Yates, J. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spec.* **5**, 976-989.

2. Perkins, D., Pappin, D., Creasy, D., Cottrell, J. (1997) Probability-based protein identification by searching sequence databases using mass-spectromety data. *Electorphoresis* **20**, 3551-3567.

3. Clauser, K. R., Baker, P., Burlingame A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem.* **71**, 2871-82.

4. Bafna, V., Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17 Suppl. 1**, S13-21.

5. Havilio, M., Haddad, Y., Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem.* **75**, 435-44.

6. Craig, R., Beavis, R. C. (2004), TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-7.

7. Geer, L. Y. , Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J Proteome Res.* **3**, 958-64.

8. Moore, R. E., Young, M. K., Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom.* **13**, 378-86.

9. MacCoss, M. J., Wu, C. C., Yates, J. R. 3rd. (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem.* **74**, 5593-9.

10. Zhang, N., Aebersold, R., Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2**, 1406-12.

11. Keller, A., Nesvizhskii, A. I., Kolker, E. Aebersold, R.. (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem*. **74,** 5383-5392.

12. Sadygov, R. G., Yates, J. R. 3rd. (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem.* **75**, 3792-8.

13. Fenyo, D., Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem*. **75**, 768-74.

14. Anderson, D. C., Li, W., Payan, D. G., Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res.* **2**, 137-46.

15. Eriksson, J, Fenyo, D. (2004) Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J Proteome Res.* **3**, 32-6.

16. Sun W, Li F, Wang J, Zheng D, Gao Y. (2004) AMASS: software for automatically validating the quality of MS/MS spectrum from SEQUEST results. *Mol Cell Proteomics*. **3**, 1194-9.

17. Washburn, M. P., Wolters, D., Yates, J. R. 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. **19**, 242-7.

18. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*. **2**, 43-50.

19. Dempster, A. P., Laird, N. M., Rubin, D. B. (1977) Maximum likelihood from incomplete data via EM algorithm. *J Roy Stat Soc Series B* **39**, 1-38.

20. Liu, J., Beaudrie, C. E. H., Yanofsky, C., Carrillo, B., Boismenu, D., Morales, F. R., Bell, A., Kearney, R. E.. A Statistical Model for Estimating Reliability of Peptide Identifications Using Mascot. ASMS 2005, Poster WP21389.

21. Hastie, T., Tibshirani, R., Friedman, J. H. (2001) *Elements of Statistical Learning.* Springer, New York.

22. Freund, Y. and Schapire, R. (1995) A decision theoretic generalization of on-line learning and an application to boosting. *Proceedings of the 2nd European Conference on Computational Learning Theory.*

23. Breiman, L. (2001) Random Forests. *Machine Learning* **45**, 5-32.

24. Vapnik V. N. (1999) *The Nature of Statistical Learning Theory.* Springer, New York.

25. Jaakkola, T., Diekhans, M. Haussler, D. (1999) Using the Fisher kernel method to detect remote protein homologies. *Proc Int Conf Intell Syst Mol Bio*, 149-158.

26. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906-914.

27. Brown M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr., Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97,** 262-7.

28. Strahler, J. R., Veine, D., Walker, A., Kachman, M., Ulintz, P., Falkner, J. (2005) A publicly available dataset of MALDI-TOF/TOF mass spectra of known proteins. *53rd ASMS Conf. Mass Spectrom. Allied Topics*, San Antonio, TX, TP22-398.

29. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*. **6**, 207-12.

30. Wysocki, V. H., Tsaprailis, G., Smith, L. L., Breci, L.A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom*. **35**, 1399-406.

31. Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S., Moritz, R. L, O'Hair, R. A., Speed, T. P., Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem*. **75**, 6251-64.

32. Tabb D. L., Huang, Y., Wysocki, V. H., Yates, J. R. 3rd. (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem*. **76**, 1243-8.

33. Platt (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In ***Advances in Large Margin Classifiers***, A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, eds., pp. 61-74, MIT Press.

34. Caruana, R., Niculescu, S., Rao, B., Simms, C. (2003) Evaluating the C-section rate of different physician practices: using machine learning to model standard practice. *The American Medical Informatics Conference (AMIA)*.

35. Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem*. **75**, 4646-58.

36. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., Yates, J. R. III (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology* **17**, 676 – 682.

37. Graumann, J., Dunipace, L. A., Seol, J. H., McDonald, W. H., Yates, J. R. III, Wold, B. J., Deshaies, R. J. (2004) Applicability of tandem affinity purification MudPIT to pathway proteomics in yeast. *Mol Cell Proteomics* **3**, 226-37.

Chapter 3

Investigating MS^2-MS^3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence[9]

3.1 Summary

Improvements in ion trap instrumentation have made n-dimensional mass spectrometry more practical. The overall goal of the study is to describe a model for making use of MS^2 and MS^3 information in mass spectrometry experiments. A statistical model is presented for adjusting peptide identification probabilities based on the combined information obtained by coupling peptide assignments of consecutive MS^2 and MS^3 spectra. Using two data sets, a mixture of known proteins and a complex phosphopeptide-enriched sample, I demonstrate an increase in discriminating power of the adjusted probabilities compared to models using MS^2 or MS^3 data only. This work

[9] The study outlined in the chapter was originally published by Ulintz PJ, Bodenmiller B, Andrews PC, Aebersold R, Nesvizhskii AI. as "Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence." Mol Cell Proteomics. 2008 Jan;7(1):71-87.

also addresses the overall value of generating MS^3 data as compared to an MS^2-only approach, with a focus on the analysis of phosphopeptide data.

3.2 Introduction

Advances in mass spectrometer design continue to propel proteomics research. One of the most widely used mass analyzers for protein work has historically been the ion trap, and a large proportion of the data from current mass spectrometry-based proteomics experiments are generated on such instruments. This trend continues with current generation 'linear trap' instruments that are characterized by increased ion capacity and thus improved resolution and sensitivity (1,2). Standard proteomics approaches are based on the predictable fragmentation of peptides in the collision cell of the mass spectrometer and the subsequent interpretation of the resulting spectra to infer amino acid sequence, referred to as tandem mass spectrometry (MS/MS or MS^2) (3-7). In practice, however, acquired MS/MS spectra are often noisy, contain only a small number of fragment ions due to incomplete peptide fragmentation, or reflect unanticipated instrumental or chemical artifacts. As a result, in a typical analysis of MS/MS spectra generated in a large scale experiment, a small fraction (frequently as low as 15%) of the spectra can be successfully interpreted and assigned a peptide sequence with high confidence (8,9).

Newer instrumentation supports alternative techniques for data generation that have the potential to improve peptide and protein identification. One such technique is 3-stage mass spectrometry (MS^3), in which peptide ions in an ion trap or ICR mass spectrometer are subjected to an additional stage of isolation and fragmentation. The faster acquisition times of newer linear trap instruments such as the LTQ provide the option of collecting MS^3 spectra of abundant MS^2 peaks with overall cycle times similar to those of normal MS/MS^2 cycles on older 3D trap instruments. As a result, a number of researchers are choosing to routinely collect MS^3 spectra during LC-MS/MS runs which have the potential to provide additional information useful for peptide identification and characterization. This is deemed particularly important in the case of proteins identified

by single peptides (10, 11) and for the analysis of phosphopeptides, the spectra of which are frequently dominated by a major fragment ion representing neutral loss of the phosphate group from the precursor peptide. Therefore, phosphopeptides have been analyzed by automated data-dependent triggering of MS^3 acquisition whenever the dominant neutral loss ion of the appropriate mass is detected in an MS^2 spectrum (12-14). Fragmentation of the neutral loss ion typically provides significantly increased structural information via increased peptide bond cleavage. Similar approaches may be applied to other major neutral loss ions (e.g. loss of 64 Da from peptides containing methionine sulfoxide) and to excessive prolyl- or aspartyl-directed fragmentation. MS^3 spectra have proven to be useful in top-down analysis as well, both for protein identification and for characterization of specific sites of post-translational modification. (15, 16)

Generally speaking, there are several ways of combining MS^2 and MS^3 spectra from the same peptide to improve peptide identification. One strategy involves integrating matching MS^2 and MS^3 spectra directly at the spectrum level, generating an "intersection spectrum" that contains only one type of ion, thus allowing simplified *de novo* sequencing of the peptide. This approach has been described by Zhang and McElvain, who demonstrated the technique's usefulness in protein sequencing (17). Olsen and Mann describe a custom scoring algorithm for MS^3 spectra: their final score for a peptide is the product of the Mascot-generated MS^2 and the custom MS^3 score (11). In glycoproteomics, it is frequently the case that MS^2 and MS^3 provide complementary structural information on a glycopeptide: information on the structure of side-chain carbohydrate moieties is generally obtained from the MS^2 spectrum, while amino acid sequence information is more readily obtained in the MS^3 (18). In the top-down technique described by Zabrouskov et al (16), sequence tags are extracted from MS^3 spectra using a *de novo* algorithm and used to complement correlated MS^2 spectral data in a "hybrid" database search strategy, implemented in the ProSight PTM search engine (19).

Related to the problem of MS^2-MS^3 spectrum integration, *de novo* sequencing-based algorithms have been described for combining pairs of spectra corresponding to unmodified and modified versions of the same peptide, or pairs of spectra corresponding to the same peptide tagged with a light or heavy version of a labeling reagent (20-23).

However, while *de novo* sequencing approaches are promising, no computational tools are currently available that can be robustly applied in a high throughput environment. As a result, analysis of MS^2 and MS^3 data is still largely carried out with a conventional database search approach using commercially available programs such as SEQUEST, MASCOT, SpectrumMill, Phenyx, Paragon, or open source programs X! Tandem, OMMSA, Inspect, or ProbID (24-29).

While all existing database search tools can be used to identify peptides from both MS^2 and MS^3 spectra, automated analysis of those different types of spectra may not be identical. This often leads to the requirement that MS^2 and MS^3 spectra be separated for processing. The main reason for this is that the measured precursor mass associated with MS^3 spectra will not always correspond to the mass of an appropriate database peptide calculated using the same conventional rules that are applied in the case of MS^2 spectra. For example, in phosphopeptide analyses variable modifications of -18 Da due to loss of phosphoric acid from S or T residues need to be specified for MS^3, while the normal +80 Da phosphorylation modification on S, T, and Y are used for MS^2. It is computationally inefficient, and an unnecessary source of false positive identifications, to perform a combined search which permits both the -18 Da loss for MS^2 spectra and the +80 Da addition for MS^3 spectra.

Searching MS^3 spectra separately from their parent MS^2 spectra essentially decouples the two sets of scans. Intuitively, if analysis of successive MS^2 and MS^3 scans results in matching peptide sequences, there is an increased confidence in both identifications. The work described here attempts to provide a general, statistically sound assessment of the confidence achieved by combining the search results of MS^2 and MS^3 spectra from the same peptide. In contrast to aforementioned work, a workflow is assumed in which the MS^2 and MS^3 spectra are searched independently using a common search engine (namely, SEQUEST in this work) and are independently statistically validated using PeptideProphet. Matching consecutive MS^2 and MS^3 scans are then re-coupled and the peptide probabilities initially computed by PeptideProphet are adjusted to account for the new "linked" MS^2-MS^3 information. A model is described that produces an adjusted probability of peptide identifications and demonstrates, using a data set of MS^2 and MS^3 spectra generated using a control protein mixture, that such a

correction can be used to better discriminate between correct and incorrect database search results. Ways to combine the adjusted MS^2 and MS^3 probabilities to compute a single confidence measure for their corresponding unique peptide are also investigated. The utility of the method is then further demonstrated using a phosphopeptide-enriched data set generated from *D. melanogaster* samples on an LTQ linear ion trap instrument. Finally, runs in which both MS^2 and MS^3 spectra are generated with an MS^2-only method are compared to address the overall benefit of generating MS^3 data.

3.3 Experimental Procedures

3.3.1 Sample Preparation and Mass Spectrometry

Two experimental data sets of MS/MS spectra were used in this work to evaluate the statistical model and to investigate its utility in the analysis of phosphopeptide-enriched samples. All spectra were acquired using an electrospray ionization (ESI) linear ion trap tandem mass spectrometer (Thermo Electron's LTQ).

(1) Nine-Protein Mix ("9-Mix") sample. A mixture of nine commercially available protein standards-- P68082, myoglobin of Equus caballus (horse); P00698 Lysozyme C precursor of Gallus gallus (Chicken); Q29443 Serotransferrin precursor (Transferrin) of Bos Taurus; P18915 Carbonic anhydrase 6 precursor of Bos taurus (Bovine); P12763, Alpha-2-HS-glycoprotein precursor (Fetuin-A) of Bos taurus (Bovine); P02754 Beta-lactoglobulin precursor (Beta-LG) of Bos taurus (Bovine); P62894 Cytochrome C of Bos taurus (Bovine); P02666 Beta-casein precursor of Bos taurus (Bovine); P02769 Serum albumin precursor (BSA) of Bos taurus (Bovine)-- was digested using trypsin and the resulting peptide mixtures were purified using reverse phase chromatography prior to mass spectrometric analysis. For the analysis of the peptides using mass spectrometry see "Mass spectrometry". The final data set consisted of three LC-MS/MS runs, with 58081 MS/MS spectra in total.

(2) "Phosphopeptide sample". This sample is a trypsin-digested, IMAC-enriched *D. melanogaster* whole cell lysate. The preparation of the phosphopeptide samples is described in detail in Bodenmiller et al. (30). Several mass spectrometry analyses of this

sample were conducted, both for analysis of performance of the probability model and to test the value of generating MS^3 data.

Mass Spectrometry

An LTQ quadrupole linear ion trap mass spectrometer (ThermoElectron, San Jose, CA) was used with a HP 1100 solvent delivery system (Agilent, Polo Alto, CA) for the analysis of the *D. melanogaster* Kc167 cells cytosolic phosphoproteome. Peptides were loaded on a capillary (BGB Analytik, Böckten, Switzerland) reverse-phase C_{18} column (75 μm i.d. and 11 cm of bed length with Magic C18 AQ 5 μm 200Å resin (Michrom BioResources, Auburn, CA, USA)), and then eluted from the capillary column at a flow rate of 200-300 nl/min to the mass spectrometer through an integrated electrospray emitter tip. Peptides were eluted for each analysis from 12% to 33% acetonitrile in which the ions were detected, isolated, and fragmented in a completely automated fashion. The exact settings for MS^n acquisition were as follows:

9 protein mix

In the first scan event, all peptides eluting from the column were recorded in MS mode. The most intense ion was selected for product ion spectrum (MS^2) in the second event. An MS^3 spectrum of the most intense peak in the MS^2 spectrum was automatically selected in the third scan event. The second and third events are then repeated two more times in the cycle, for the second and third most abundant MS^1 ions, for a total cycle of seven events. A threshold of 5,000 ion counts was used for triggering an MS^2 attempt. Wideband activation was enabled for all MS^2 and MS^3 scan events; MS^2 isolation width was set to 2.0 m/z and MS^3 isolation width was set to 4 m/z. For triggering an MS^3 event the most intense ion had to be above 50 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

Phosphopeptide sample

All peptides eluting from the column were recorded in MS mode in the first scan event. The most intense ion was selected for product ion spectrum (MS^2) in the second event. An MS^3 spectrum of the most intense peak in the MS^2 spectrum, which for the phosphopeptide containing sample is in most cases the neutral loss peak (of 98 Da) from

a serine/threonine phosphopeptide, was automatically selected in the third scan event. These three events form one complete cycle. A threshold of 20,000 ion counts was used for triggering an MS^2 attempt. Wideband activation was enabled for all MS^2 and MS^3 scan events. MS^2 isolation width was set to 2 m/z and MS^3 isolation width was set to 3 m/z. For triggering an MS^3 event the most intense ion had to be above 500 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

Phosphopeptide sample – additional data sets for comparison of MS^2-only with MS^2/MS^3 methods

For the MS^2/MS^3 data set the data-dependent MS^n spectra were acquired as follows: in the first scan event, all peptides eluting from the column were recorded in MS mode, and then the most intense ion was selected for product ion spectrum (MS^2) in the second event. In the third event a MS^3 spectrum was triggered specifically in the event of a phosphate neutral loss (-98 Da for singly, -49 Da for doubly and -32.66 Da for triply charged peptides) in the MS^2 event. The second and third events are then repeated two more times in the cycle, for the second and third most abundant MS^1 ions, for a total cycle of seven events. For the MS^2-only data set the data-dependent MS^n spectra were acquired as follows: in the first scan event, all peptides eluting from the column were recorded in MS mode, and then the three most intense ions were consecutively selected for product ion spectrum (MS^2) for a total cycle of four events. Further settings for these samples were: wideband activation was enabled for all MS^2 and MS^3 scan events, MS^2 isolation width was set to 2 m/z and MS^3 isolation width was set to 4 m/z. For triggering an MS^3 event in the MS^2/MS^3 data set the most intense ion had to be above 50 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

3.3.2 Database Searching and Results Analysis

MzXML files were generated from ThermoFinnigan *.raw files using the ReAdW tool available in the TPP platform (31-33). MS^2 and MS^3 peaklist files in *.dta format were extracted separately from the mzXML files using mzXML2Other tool with the -

level option[10]. For the 9-Mix data set, a custom fasta sequence file was constructed consisting of sequences corresponding to the proteins in the mixture and common contaminants appended to a reversed version of the IPI Human data set. Resulting *.dta files for the 9-Mix data set were searched with SEQUEST using the following parameters: peptide tolerance of 3.0 Da; b- and y-ion series; partial trypsin digestion, allowing for one missed cleavage site; a fixed modification of 57.02 was specified for Cysteine and a variable post-translational modification (PTM) of 16.0 to Methionine. MS^3 data sets were searched using identical parameters. Note, partial trypsin specificity is required for searching MS^3 spectra corresponding to the fragmentation of a selected y- or b-ion from the MS^2 spectrum. If sufficient computational resources are available, searching MS^2 spectra allowing for partially tryptic peptides can often be beneficial and result in additional identifications. However, doing so requires that the results are properly analyzed with a tool that accommodates tryptic termini information in the statistical model, such as PeptideProphet. In addition, a subset of MS^3 spectra from this data set was also searched allowing for the C-terminus variable modification of -18.0 Da to accommodate the possibility that the MS^3 precursor is a b-ion (11). The results indicated that including this modification does not significantly alter the overall performance; in fact, accommodating the variable modification decreases the number of identifications slightly (due to loss of a number of true peptide assignments because of increases in search space). Based on this, the C-terminal modification was not used in the final analysis of data presented in this chapter. The resulting data set contained 76873 peptide assignments, counting 2+/3+ duplicates: 48921 MS^2 (554 singly charged, 24233 doubly charged, and 24134 triply charged), and 27952 MS^3 (4582, 11700, and 11670 singly, doubly, and triply charged, respectively). Note that because of the charge state ambiguity (in the case of low mass accuracy data such as the data sets used in this work, the charge state of a multiple charged peptide ion cannot be reliable determined), most of the multiply charged spectra were searched twice, assuming 2+ or 3+ charge state. Furthermore, due to a relatively small number of singly charged MS^2 spectra, all such spectra were left out of the subsequent analysis.

[10] http://tools.proteomecenter.org/software.php

The database for the phosphopeptide-enriched samples consisted of all *Drosophila melanogaster* sequences exported from the UniProt database (34), 26311 entries total, to which the reversed set of sequences was appended. Parameters for the MS^2 search were: peptide tolerance of 3.0 Da; partial trypsin digestion, one possible missed cleavage; fixed modification of 57.02 for Cysteine; variable modifications of 80 Da were specified for S, T, and Y; a maximum 4 PTMs per peptide. The MS^3 spectra were searched with the same set of parameters except that variable modifications of -18 Da on S and T (instead of +80 Da) were specified to accommodate loss of phosphoric acid leading to a dehydroalanine or dehydrobuyric acid, respectively. SEQUEST database searching for the primary phosphopeptide data set (excluding the MS^2/MS^3 to MS^2-only comparisons) resulted in 28865 peptide assignments, counting 2+/3+ duplicates: 16647 MS^2 (143 singly charged, 8483 doubly charged, and 8021 triply charged), and 12218 MS^3 (547, 5895, and 5776 singly, doubly, and triply charged, respectively).

The additional phosphopeptide-enriched data sets used for comparison of MS^2/MS^3 and MS^2-only methodologies consisted of the following number of peptide assignments following SEQUEST database searching- Run 1 (A07_5205): 4915 MS^2 assignments (95 singly charged, and 2410 each of doubly and charged), and 1897 MS^3 assignments (31 singly and 933 each doubly and triply charged); Run2 (A07_5206): 6450 MS^2 assignments (126 singly, 3162 doubly and triply charged); Run 3 (A07_5207): 4883 MS^2 assignments (103 singly charged, and 2390 each of doubly and charged), and 1879 MS^3 assignments (43 singly and 918 doubly and triply charged); and Run 4 (A07_5208): 6403 MS^2 assignments (159 singly, 3122 doubly and triply charged).

3.3.3 Processing of MS^2 and MS^3 search results

Search results for each LC-MS/MS run were generated by first producing an html results file using the out2summary tool, exporting one result file for each MS level, for each run: a total of six files for the 9-Mix data set and two files for the phospho data set. Html results were then converted into pepXML format (31) using Sequest2XML. PeptideProphet (32) was run on each result set, generating probability scores for each search result that are added to the pepXML documents.. For the phospho data sets,

PeptideProphet was run with the "-l" option, which results in alternate processing of DeltaCn scores marked with '*', results for which the top and second-highest ranked peptide assignment to a spectrum have homologous sequences (>70% sequence identity). With this option on, PeptideProphet will use the Xcorr score of the first non-homologous lower scoring peptide match when computing DeltaCn score of the best scoring peptide. This option is beneficial in the event that the search returns several identical results that differ only by modification site for a sequence, as often occurs in phoshorylated peptide identifications.[11] Resulting files were parsed and processed to generate all matching statistics using a custom set of scripts implemented in Python. Certain subsets of data were also exported into a local Mysql database instance to facilitate generation of specific statistics.

3.3.4 Linking MS^2 and MS^3 scans and search results

The spectra in these experiments were generated in an interlaced manner, i.e. the scan cycle on the instrument followed the format: MS^1->MS^2->MS^3->MS^2->MS^3->MS^2->MS^3, or MS^1->MS^2->MS^3, with the MS^2 scans triggered in a data dependent manner from the MS^1, and the MS^3 scans triggered from the preceding MS^2. As a result, a set of "linked" MS^2/MS^3 scans were generated based on consecutive scan numbers. In the resulting data set, MS^2 scans with no consecutive MS^3 were retained and designated as linked, but as a link to a null MS^3 identification. MS^3 scans without preceding MS^2 scans should not occur physically, but do in these data for several reasons: namely, the corresponding MS^2 peaklists that produced no database search result are typically not reported. Also, some spectra containing only a few peaks may be filtered out by the data conversion software. The small number of instances in which these "orphaned" MS^3 scans are generated invariably result in incorrect peptide identifications and are eliminated from subsequent analysis.

Due to uncertainty with the charge state each multiply charged scan was searched twice (in both 2+ and 3+ charge state), resulting in multiple search results for each scan.

[11] The default option in PeptideProphet is to set SEQUEST DeltaCn score to zero to reduce the probability that the best scoring peptide assignment to a spectrum is correct when the second best scoring peptide has high sequence homology.

Consideration needs to be given to potential links between MS^2 and MS^3 search results for any pair of scan numbers. A +1 MS^2 search result may only be linked to an MS^3 search result that is +1, and a +2 MS^2 scan may produce a link to a search result with either a +1 or +2 charge state. The double- and triple-charged SEQUEST search duplication, however, creates a situation in which a +3 MS^2 search result may produce two possible links to +2 and +3 MS^3 search results for any pair of scan numbers. After generating all possible links, one pair of search results amongst all possible pairs for any two scan numbers (designated as the "unique pair") is selected based on whether the sequences of the two peptide identifications composing a pair are matching. Matching is defined here as whether or not the sequences are equal, or whether one contains a subsequence of the other. For non-matching pairs, and scan sets with more than one pair with matching sequences, the match pair with the highest summed PeptideProphet probability is designated as the unique pair. A schematic of all matching possibilities and selection of a unique pair is shown in Figure 3-1.

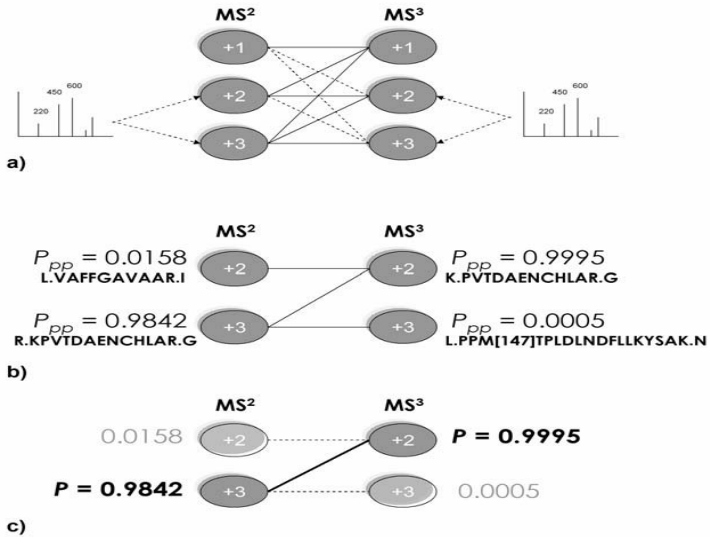


Figure 3-1. MS^2 and MS^3 charge state linking possibilities and unique pair selection. Panel a) indicates all possible linked pair possibilities. Solid lines indicate physically possible link pairs, dashed lines between charge states indicate transitions that should not occur. The redundant +2 and +3 search results generated for an individual spectrum to accommodate charge ambiguity are schematically represented. b) An example set of three linked pairs for two consecutive scan numbers. The top-scoring sequences and the corresponding PeptideProphet probabilities are shown. c) The unique pair that is selected for the two MS levels is indicated with a bold line.

3.4 Results and Discussion

3.4.1 Overview of the probability adjustment method

The overall methodology for the approach is outlined in Figure 3-2. Data generated by the mass spectrometer are processed via the Trans-Proteomic Pipeline (TPP) following normal procedures and using SEQUEST, Mascot, or X! Tandem database search tools for peptide identification (the tools currently supported by the TPP), up through generation of peptide probabilities from PeptideProphet (32). Analyses in this early stage of processing are conducted separately for MS^2 and MS^3 data. To calculate an adjusted probability for all assignments, successive scans must be linked as described in the Methods section. The multiple potential matches resulting from the charge state ambiguity are reduced in the processing, retaining only the most probable matching pair for any two scan numbers.

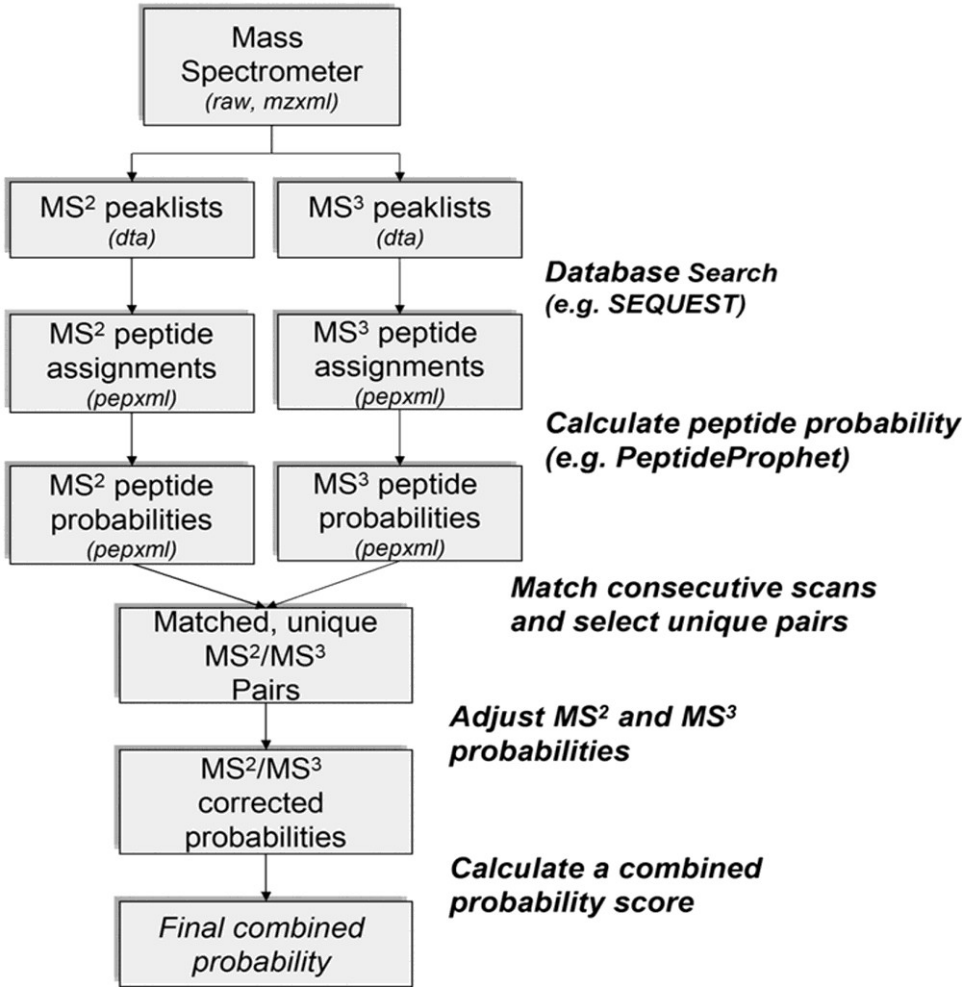


Figure 3-2. Overview of methodology. MS^2 and MS^3 spectra are extracted from the raw data and the spectra are assigned peptides using sequence database searching (SEQUEST or similar programs). The resulting peptide assignments are statistically validated using PeptideProphet, which calculates for each assignment in the data set a probability of being correct (applied separately for MS^2 and MS^3 data). MS^2 and MS3 scan results are correlated based on scan number, in which an MS^3 spectrum is linked to an MS^2 if its scan number is consecutive. Based on the overall matched data set, a Bayesian probability correction is applied to linked scan results individually for MS^2 and MS^3 spectra, resulting in adjusted probability scores. In the final step, the MS^2 and MS^3 scan results are combined and a final probability calculated for each scan number as representative of the peptide identification.

Based on the sequence of the highest scoring peptide produced by the database search tool for each scan, consecutive MS^2/MS^3 pairs may then be classified as to whether or not they match the same peptide sequence. This classification forms the basis

for the adjusted probability score (see below), which functions to reward assignments with matching sequences. Only the top-ranked peptide sequence for each spectrum is used in this analysis; accommodation of lower ranking results, while potentially useful, is not considered for simplicity. The result of the probability correction procedure is a data set of linked MS^2 and MS^3 peptide identifications with adjusted probability scores.

3.4.2 Linking MS^2 and MS^3 data: a case study of the 9-Mix data set

This analysis is carried out using a mixture of purified proteins (9-protein mix data set), in which it is possible to confidently label peptide identifications as 'correct' or 'incorrect'. Because this data set was searched against a database consisting of the sequences of the mixture proteins appended with a much larger reversed human protein sequence database, each spectrum could be assigned a correctness label based on whether the top SEQUEST hit for the spectrum was to one of the known protein entries. The method used was simply to label as incorrect any assignment of a peptide from a known incorrect database entry (reversed human protein sequence entries in this case), whereas all assignments of peptides to one of the sample proteins can be considered correct (32).

The procedure begins by linking consecutive MS^2 and MS^3 scans using their scan numbers. Overall, there were 48921 MS^2 spectra and 27952 MS^3 spectra generated for the 9-Mix dataset. Due to the uncertainty in the precursor charge state for LTQ spectra, many spectra are redundant; for any pair of consecutive MS^2/MS^3 scan numbers, there may be one or two SEQUEST search results generated for each MS level, as described in Experimental Procedures. Consequently, an MS^2 search result may be linked to more than one MS^3 search result. For the 9-Mix data set, there are 16140 unique linked pairs in which the MS^3 is not null. Amongst these, eighty nine have MS^2/MS^3 charge states of +1/+1, eight of which match "correct" protein sequences in the database (either one or both of the sequences match). For doubly-charged MS^2 pairs, 3761 are +2/+2 and 4043 are +2/+1, of which 878 and 2020 are correct, respectively. For triply-charged MS^2, +3/+3: 4020 pairs, 631 correct; +3/+2: 3777 pairs, 1177 correct; and +3/+1: 450 pairs, 111 of which are correct. In all, linked pairs in which the MS^3 has one less charge than

the MS^2 are more likely to be correct. However, linked pairs for which the MS^3 is the same charge state as MS^2 account for 36% of the correct identifications.

Neutral loss of amino acids from the N- and C-termini is a common phenomenon and has been described previously (35, 36). Selecting linked pairs in which both MS^2 and MS^3 sequences are labeled correct and of the same charge state (+1/+1, +2/+2, and +3/+3) allows us to identify examples of amino acid neutral loss. Our data confirms the conventional rules for amino acid neutral loss described in the literature. Virtually all examples correspond to N-terminal loss of 1-4 amino acid residues, most frequently N-terminal to a proline. 276 out of 323 of the occurrences are doubly-charged, three are singly-charged, and the remaining forty-four triply-charged. Most examples occur multiple times: in all there are one, thirty-four, and nine unique neutral loss sequence examples for the singly-, doubly-, and triply-charged cases, respectively. These examples are provided in Supplementary Table S1.

After linking consecutive scans and selecting a unique linked pair, the peptide assignments are binned into sequence match categories dependent on whether a consecutive scan exists, and if so, whether the top-scoring SEQUEST sequence result of the successive scans match (Table 3-1). Sequence match categories (referred to as Match categories, or simply 'Match' later in the text) are defined as follows: 0) no consecutive scan; 1) consecutive scans, but MS^2 and MS^3 sequences do not match; 2) consecutive scans, MS^3 sequence is a subset of the MS^2 sequence; 3) consecutive scans, MS^3 sequence identical to the MS^2 sequence; and 4) consecutive scans, MS^2 sequence a subset of MS^3 sequence. In the data set of unique pairs, 69% of all MS^2 spectra produced consecutive MS^3 spectra (16140). Out of those consecutive pairs, 1458 (9%) had matching sequences in which the MS^3 sequence was a subset of the MS^2 sequence. 116 MS^3 spectra were orphaned because they did not have a preceding MS^2 scan, and were discounted. I note that there were no instances of identical sequence matches between MS^2 and MS^3 top-scoring hits in the 9-Mix data set, as may occur for neutral-ion events in which only a side-chain moiety is lost from the otherwise intact peptide backbone (e.g. a phosphate). These losses are observed in other similar data sets, however, and do occur in the phospho-enriched data sets described later.

	0: No Consecutive	**1**:Consecutive (no match)	**2**: MS^3 Seq In MS^2 Seq	**3**:MS^3 Seq = MS^2 Seq	**4**: MS^2 Seq In MS^3 Seq
Unique Matches	**7227**	**14665**	**1458**	**0**	**17**

Table 3-1. Results of binning consecutive MS^2/MS^3 scan pairs for the 9-Mix data set into sequence match categories. Counts indicate the number of unique pairs as described in the text. Seq, Sequence.

For a small number of linked pairs, the top-scoring MS^3 sequence appears to be a superset of the MS^2 sequence, binned as sequence match category 4. Clearly such pairs are not physically possible. Detailed analysis indicated that that most of those cases can be explained as resulting from misidentification of the true peptide sequence from either MS^2 or MS^3 scan. For example, in some of these instances, the sequence corresponding to the +2 MS^2 is a subsequence of both the +3 MS^2 sequence and the +2 MS^3 sequence, with the +2/+2 MS^2/MS^3 pair selected as the unique pair. In those cases, the peptide assignment to the +3 MS^3 peaklist (with +3 being the true charge state of the peptide ion) scored lower than the assignment of a shorter peptide (a subsequence of the true peptide) to the +2 MS^3 peaklist. Other examples involved cases of a high scoring assignment of a longer partially tryptic peptide sequence when the true peptide was a post-translationally modified tryptic peptide missed due to the restricted nature of the database search. Similarly, several cases were observed where an MS^3 scan acquired on a doubly charged b-ion fragment from the parent MS^2 spectrum resulted in a match of a longer sequence to the +3 MS^3 peaklist, and no match in the case of the correct +2 charge state. In any event, as can be seen from Table 3-1, match category 4 represents a small number of special case instances. For simplicity of articulation, this category is dropped from subsequent analysis.

Using the labeling of the data, the accuracies and sensitivities of the probability calculations could be determined. Towards this end, each linked pair of spectra can also be assigned a truth category based on the correctness of the peptide assignments to the MS^2 and MS^3 scans. The truth category is a label indicating whether neither, both, or

which one of the matching scans has a 'correct' label. The total numbers of scans in each truth category are shown in Table 3-2. The number of unique pairs of search results in which both sequences were correctly assigned is 1509, corresponding to 6.4% of the total number of unique pairs of scans. A greater number of linked pairs (3316 total, 14.2%) have either the MS^2 only assigned correctly (2029), or only the MS^3 (1287).

	+/null	-/null	+/+	+/-	-/+	-/-
Unique Matches	**1025**	**6202**	**1509**	**2029**	**1287**	**11315**

Table 3-2. Classification of consectutive MS^2/MS^3 scan pairs for the 9-Mix dataset into truth categories. A "+" in the truth category column descriptors indicate a correct match, "-" indicates an incorrect match, and "null" indicates the lack of consecutive MS^3 for an MS^2 scan.

When comparing the counts in the sequence match category bins (Table 3-1) with the truth category bins (Table 3-2), there appear to be several (thirty four) more +/+ truth matches than expected from the number of entries in the sequence match bin categories 2 and 4. These entries are the result of sequence match category 1 entries contributing to the +/+ truth bin. There are a number of cases in which the top-scoring MS^2 and MS^3 sequences both match one of the sample mix proteins, but the proteins are different or the match is to different peptides from the same protein. Most of the instances are examples of the latter case: a homologous sequence in the protein TRFE_BOVIN results in two different peptides (CLMEGAGDVAFVK and KGDVAFVK) being identified in the joined pairs. One of the commercially obtained proteins in the mixture, TRFE_BOVIN, was also contaminated with the homologous TRFL_BOVIN, which exhibits 59% sequence identity. As a result, homologous but not identical peptide sequences between the two proteins are identified in the joined pairs. For four cases, however, even though both MS^2 and MS^3 identifications in the pair are labeled correct in that individually their sequences match one of the sample proteins, there is no similarity between the matching sequences. These can be considered as chance matches to one of the sample mix proteins incorrectly labeled as correct (the observed number of such chance matches is consistent

with the expected number given the relative sizes of the 9-Mix and the reversed Human protein sequence database). In all of such cases, either the MS^2 or the MS^3 was a high-probability result with the other joined probability very low.

3.4.3 Probability adjustment calculation

In automated analysis of mass spectrometry data, one of the most important tasks is the calculation of accurate and discriminative confidence measures for each peptide assignment to a spectrum produced by a database search tool. Towards that end, we seek to calculate a correction to the probability score that accommodates the increase in confidence resulting from matching MS^2 and MS^3 spectra. The fact that matched consecutive MS^2 and MS^3 spectra are more likely to be correct forms the basis for adjusting the probabilities of these spectra.

Calculation of probabilities for each peptide assignment in the data set, performed independently for MS^2 and MS^3 data, represents the starting point in this analysis. PeptideProphet computes a probability for a peptide, designated here as $p(+|D)$, by using the mixture model EM algorithm to model the distributions of various discriminant spectrum-level parameters, collectively represented here as D. The spectrum-level information D typically includes the discriminant database search score (a linear combination of the renormalized search scores reported by the database search tool used), the number of termini consistent with the specificity of the enzyme used to digest proteins, the number of missed internal cleavage sites, and the difference between the measured and the calculated precursor ion mass. In certain cases, additional parameters are included in the model such as the peptide pI value (37), or the presence of certain residues or sequence motifs in the sequence of the assigned peptide (e.g., the presence of a cysteine in the case of ICAT experiments, or NxS/T motif in the case of experiments employing glycopeptide-enrichment strategies). PeptideProphet probabilities are reasonably accurate for both MS^2 and MS^3 spectra. A plot displaying probability accuracies of PeptideProphet results for the 9-Mix data is provided in Figure 3-3.

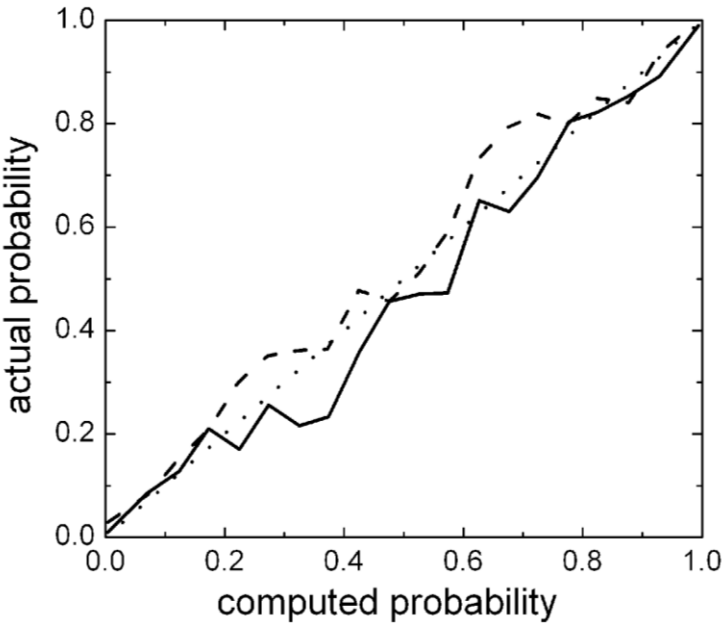


Figure 3-3. **Accuracy of PeptideProphet probability calculation for MS^2 and MS^3 identifications for the 9-Mix dat set**. MS^2 results are shown as a solid line and MS^3 as dashed. A perfect model would produce a 45° line, plotted as a dotted line. The amount of deviation of the calculated probability score from the true probability is indicated by deviation from 45°. To generate this figure, all peptide assignments were first sorted based on the calculated probability. A sliding window with a size of fifty was then applied to the ranked list, calculating the fraction of correct assignments within the window. The PeptideProphet probabilities of spectra within the window were also summed and averaged.

The approach used to accommodate the additional sequence matching information is similar to the method described in (33) for adjusting probabilities to account for additional protein level information using the number of sibling peptides (NSP). The MS^2/MS^3 sequence match information is not available at the initial data analysis step, but can be used to adjust the initial probabilities *p(+|D)* after linking the corresponding MS^2 and MS^3 scans. Again, the adjustment is performed separately for MS^2 and MS^3 level data. Given the sequence match category (Match) assignments for all linked spectra, the adjusted probability of a linked peptide assignment from a certain sequence match category, *p(+|D, Match)*, may be calculated as:

$$p(+|D,Match) = \frac{p(+|D)p(Match|+)}{p(+|D)p(Match|+) + p(-|D)p(Match|-)} \quad (1)$$

where $p(Match|+)$ and $p(Match|-)$ represent the empirically derived probabilities of observing a peptide assignment in each Match category among all (MS^2 or MS^3) correct and incorrect peptide assignments in the data set, respectively. Note that this calculation assumes that the information derived from linking consecutive scans is independent of the identification information generated by a search engine. This is largely true. Normalized PeptideProphet SEQUEST discriminant score distributions for correct and incorrect peptide assignments to MS^2 spectra of doubly charged precursor ions, plotted separately for peptide assignments to MS^2 spectra belonging to different Match categories, are shown in Figure 3-4; score distributions are similar for all values of Match parameter, justifying the assumption of the independence between the discriminant database search score and Match parameter.

The probability distribution $p(Match|+)$ may be calculated for each match category k as follows:

$$p(Match \mid +) = \frac{1}{Np(+)} \sum_{\{i|Match_i \in k\}} p(+ \mid D_i, Match_i) \tag{2}$$

where N is the total number of (MS^2 or MS^3) peptide assignments in the data set, and the sum is over all peptides i in each Match category. The term $p(Match|-)$ is calculated in a similar way. The overall proportion $p(+)$ of correct assignment in the data set may be calculated as:

$$p(+) = \frac{1}{N} \sum_{i} p(+ \mid D_i, Match_i) \tag{3}$$

The probabilities in Eq. 1, and the Match parameter distributions in Eq. 2, can be determined by starting with the initial PeptideProphet probability for each assignments, $p(+|D_i)$ and the overall proportion, $p(+)$. The probabilities and Match distributions can then be updated in an iterative manner. However, a single iteration was deemed to be sufficient for the data set used in this work.

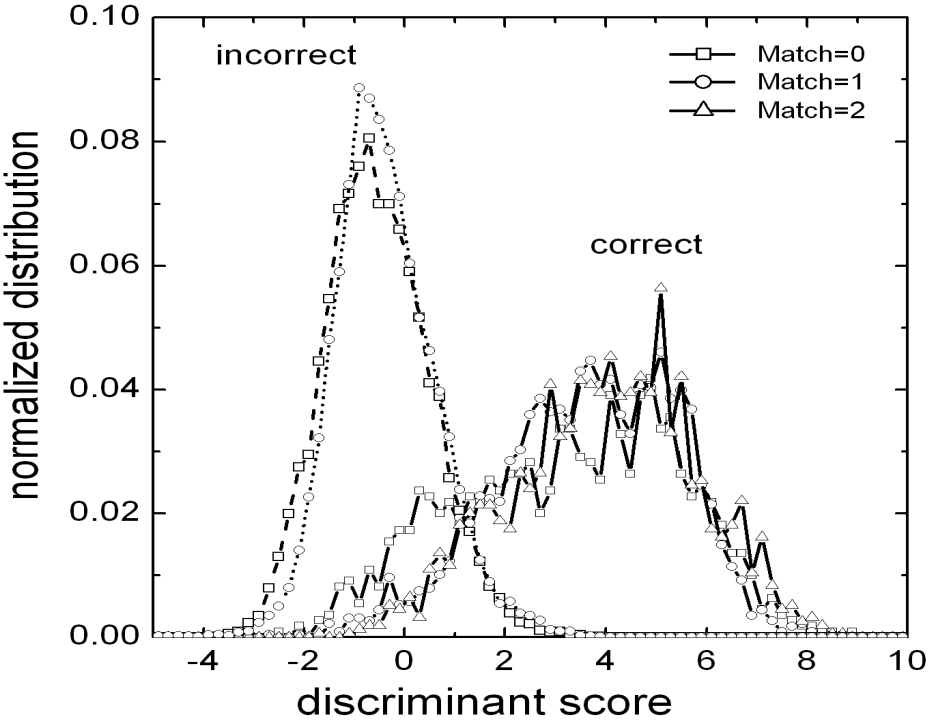


Figure 3-4. **Distributions of SEQUEST discriminant database search scores and MS^2-MS^3 match parameters**. Normalized discriminant score distribution among correct and incorrect peptide assignments to 2+ charged MS^2 spectra from the 9-Mix data set are plotted separately for peptides in Match category 0, 1, and 2.

3.4.4 Application of the probability adjustment method to the 9-Mix data set

Table 3-3 lists $p(Match|+)$ and $p(Match|-)$ distributions calculated using Eq. 2 for the 9-Mix data set for both MS^2 and MS^3 scans. It can be seen that, in the case of MS^2 spectra, a larger fraction of incorrect assignments have no consecutive matching scan. For all instances, the most likely sequence match category is category 1, corresponding to the case in which consecutive scans occur but with no matching sequence. This is perhaps intuitive in the sense that it might frequently be the case that either the MS^2 or the MS^3 will produce an identifiable sequence, but not both. The most obvious discriminating measure is the fact that for 30% of the correctly assigned MS^2 spectra (the top row in the table), the linked MS^3 spectrum was assigned a peptide sequence that is a subset of the MS^2 sequence, as opposed to a 5% incidence for incorrect MS^2

identifications. If sequence matches are observed, identifications are thus much more likely to be correct; the same argument applies for MS^3 scans preceded by MS^2 scans. Also noteworthy is the fact that for match category 1 pairs, the probability of a correct identification is less than the probability of an incorrect identification. This will result in a probability penalty for consecutively linked scans without matching sequences. The penalty is small in this case, much smaller than the boost due to a consecutive matching scan, but is nevertheless an effect of the model.

	Sequence Match Category			
	0	**1**	**2**	**3**
MS^2 *p(Match|+)*	0.213	0.483	0.304	0.000
MS^2 *p(Match|-)*	0.332	0.662	0.005	0.000
MS^3 *p(Match|+)*	0.000	0.584	0.416	0.000
MS^3 *p(Match|-)*	0.000	0.960	0.040	0.000

Table 3-3. **Posterior probabilities of observing a correctly (+) or incorrectly (-) matching peptide to a MS^2 or MS^3 scan**. Shown are results for the four most frequently observed sequence match categories in the 9-Mix data set: 0, no consecutive scan; 1, consecutive scan, no matching sequence; 2, consecutive scan, MS^3 sequence is a subset of MS^2 sequence; 3, consecutive scan, MS^3 sequence identical to MS^2 sequence.

It should be noted that in addition to classifying peptide match pairs into bins as a function of sequence matching, they can also be classified into various precursor charge state pairs. Significant differences exist between the precursor charge state distributions of correct and incorrect matches. An expansion of the sequence match category probabilities into charge category bins is provided in Figure 3-5 (Panels b-e) for each of the four posterior Match probability distributions of Table 3-3, as well as total counts of the number of matches falling into each bin for the 9-Mix data set (Panel a). The charge state information would likely provide additional discriminative power. However, further sub-classification of the data into charge state pairs requires larger amount of data and complicates the model. Thus, the charge state information has not been utilized in the model at this time.

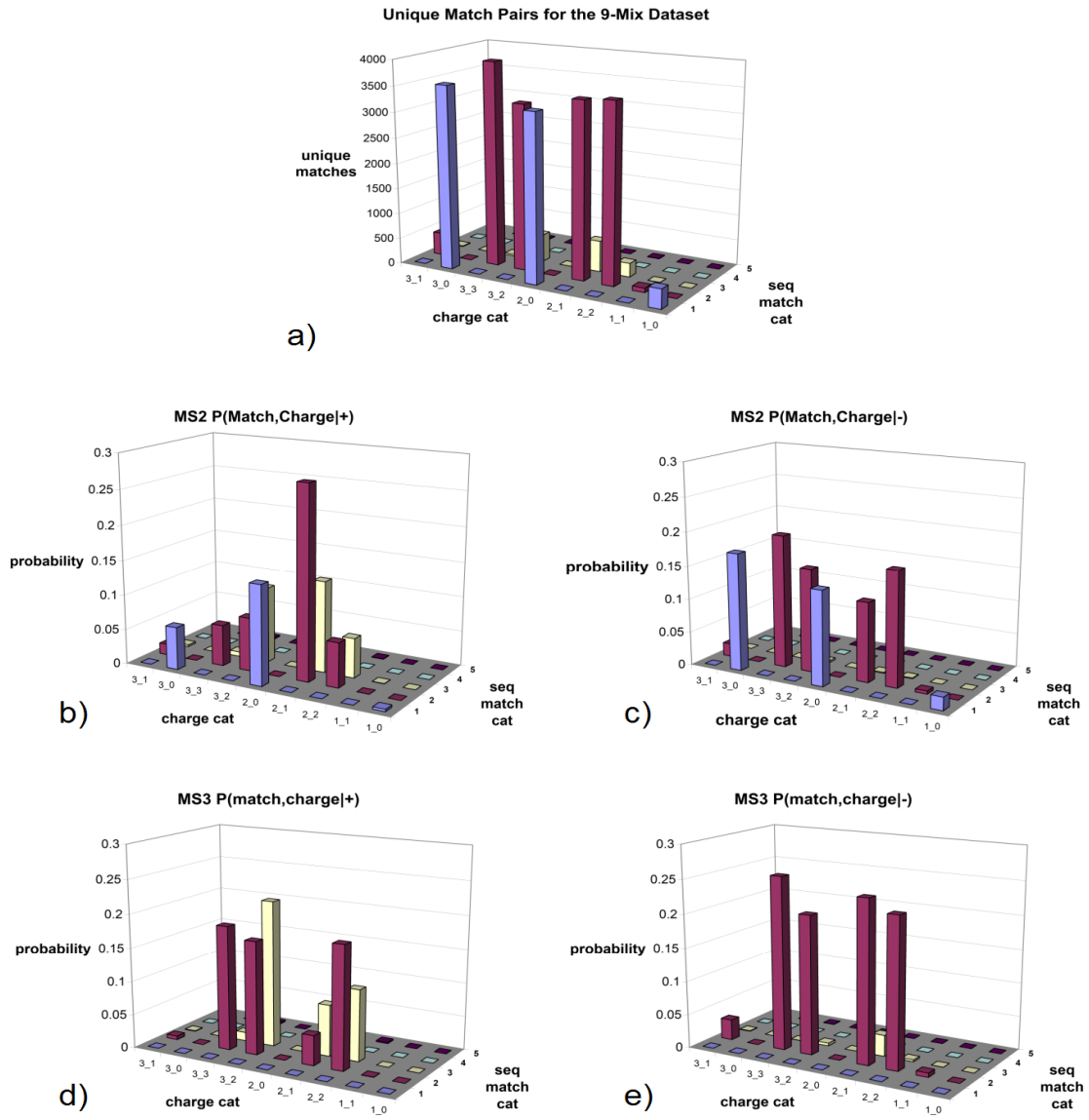


Figure 3-5. Total bin counts and posterior match probability distributions of unique matching pairs for the 9-Mix data set. Charge categories a labeled based on the charge of both parent ions; e.g the charge category "2_1" indicates a +2 MS^2 and a +1 MS^3. The "0" in the "3_0", "2_0" and "1_0" categories denotes that there is no consecutive MS^3 scan for the spectrum. Please see the main text for a description of the five sequence match categories. Panel a) shows total match pair counts for each charge/seq match bin for the 9-Mix data set. Panels b)-e) plot joint posterior probability values for each of the respective charge/seq match bins. Each of the four distributions- MS^2 correct (+), MS^2 incorrect (-), MS^3 correct, and MS^3 incorrect -are plotted separately.

An example of the probability adjustment procedure described above is illustrated in Figure 3-6 using a pair of matching scans from the 9-Mix data set. MS^2 spectrum A06_7233_c.18651.18651 is first paired to MS^3 spectrum A06_7233_c_18652.18652 by consecutive scan number. MS^2 assigned peptide sequence TLNFNAEGEPELLMLANWRPAQPLK is then compared to MS^3 sequence GEPELLMLANWRPAQPLK. Since the MS^3 sequence represents a fragment of the MS^2 sequence, the linked pair is assigned to sequence match category 2. The adjusted probabilities are then calculated for each spectrum using Eq 1. In this instance, the initial PeptideProphet probability of 0.712 is adjusted to 0.995 for the MS^2 spectrum, and 0.832 to 0.989 for the MS^3. A combined probability may then optionally be calculated for the linked pair as a new discriminating measure, as discussed later in the text.

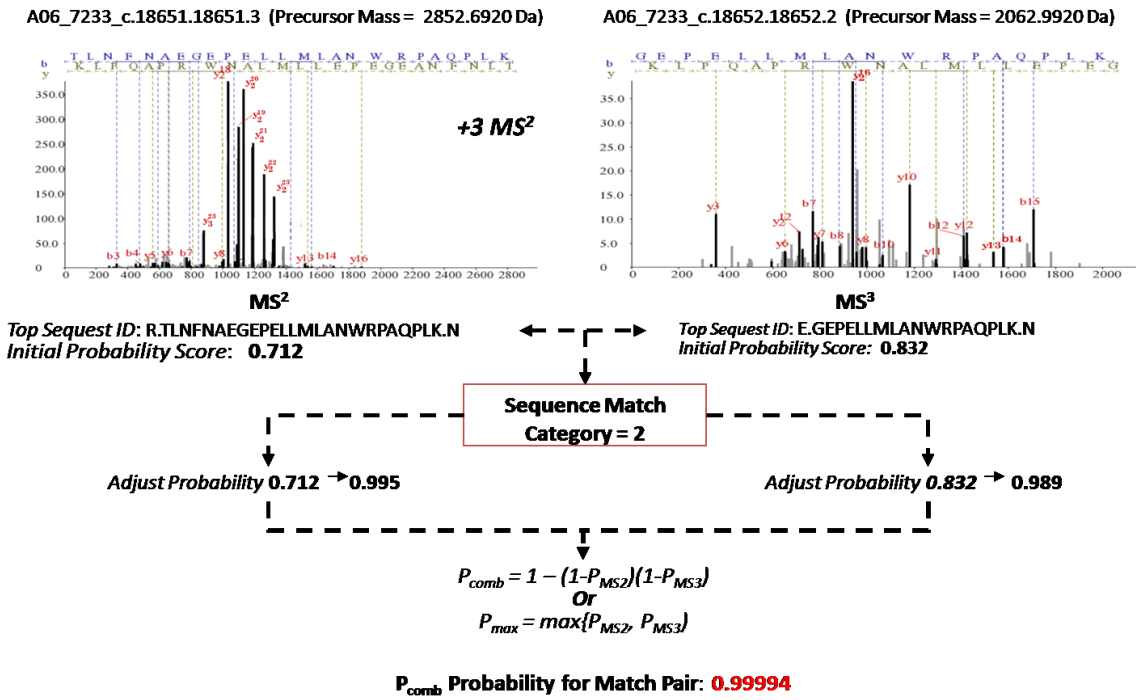


Figure 3-6. Example of MS^2/MS^3 linked pairs and the probability correction procedure. MS^2 (left) and the matching consecutive MS^3 peaklist (right) are shown. The MS^2 spectrum is triply-charged and the MS^3 spectrum doubly-charged. The sequence matches categorize the pair into Sequence Match Category 2, allowing for a probability adjustment correspondent with that category. The individual MS^2 and MS^3 probabilities may then be combined as indicated.

Indicated in Figure 3-7 are examples of fragmentation patterns from other charge state pairs. These examples are provided here to illustrate both differences in the relative extent of fragmentation that can occur as a function of charge and also the presence of redundant ions appearing in both the MS^2 and MS^3 spectra. Panels 3-7a – 3-7c contain examples from the phospho data set, specific features of which will be discussed in more detail later in the chapter. It should be noted that many identical ions can be observed between matching MS^2 and MS^3 spectra.

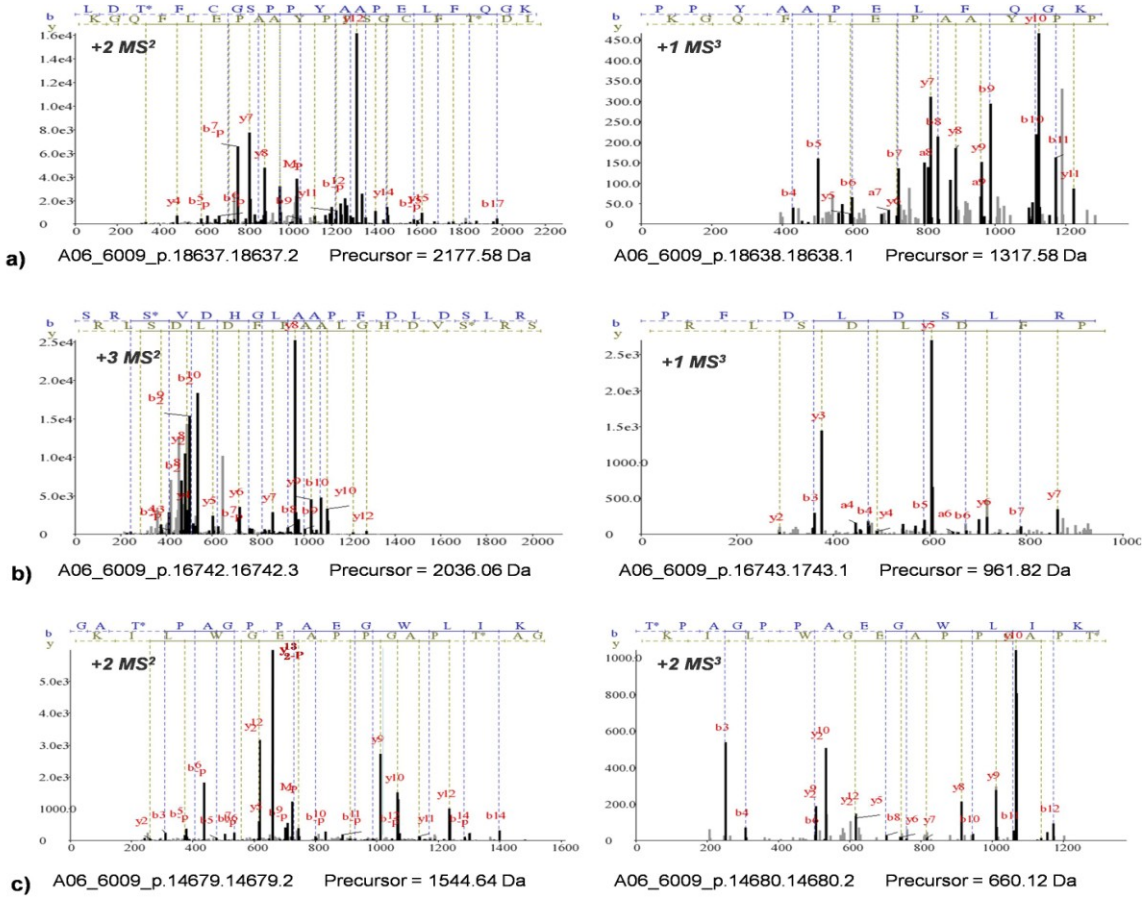


Figure 3-7. Examples of MS^2/MS^3 linked pairs with different charge states. MS^2 (left) and the matching consecutive MS^3 peaklist (right) are shown. a) A +2/+1 match pair for a phosphopeptide identification in which the y_{12} ion is selected for MS^3. b) A +3/+1 phosphopeptide identification; the y_8 ion is selected for MS^3. c) An example of a +2/+2 loss of the phosphate moiety in which the most abundant MS^2 peak selected for MS^3 is the doubly-charged y_{13} – 98 Da.

In the development of the model, several (match category 2) cases were observed where both paired spectra had a low initial probability of being correct, but their probabilities became intermediate or even high values after adjustment. For example, the initial probabilities for peptide assignments to linked scans A06_7232_c.4362.4362.3 (MS^2 scan), and A06_7231_c.4363.4363.2 (MS^3 scan) of 0.077 and 0.319 would get boosted to 0.827 and 0.830, respectively, if the probabilities were adjusted using the Match parameter distributions shown in Table 3-2. Boosting such low probability assignments may be undesirable regardless of their match category. To address this, several approaches were investigated, including introduction of probability-dependent match categories. A very simple constraint that worked well in the case of the 9-Mix data set was to avoid any probability adjustment for Category 2 matches if both initial MS^2 and MS^3 probabilities were below a specified threshold, 0.5 in the case of these data. This was an optional feature that was investigated using the 9-Mix data set but not utilized for the phosphopeptide data sets, as it was deemed a minor adjustment that did not significantly affect the overall results; specifically, the number of entries in the 9-mix data set that were affected by this exception was only 24 out of a total 23367 unique matches

The improved discriminatory power of the adjusted probabilities, calculated using the $p(Match|+)$ and $p(Match|-)$ distributions shown in Table 3-3 (after the empirical correction described above), is indicated in Figure 3-8, which shows Receiver-Operator Curves (ROC) for the data. The performance of the model is evaluated separately for MS^2 and MS^3 spectra. The false positive error rate is plotted as a function of the sensitivity attainable by selecting a variable probability threshold. Sensitivity in this case is defined as the ratio of the number of correct peptide assignments to MS^2 (Figure 3-8a) or MS^3 scans (Figure 3-8b) with a probability greater than or equal to a specific probability threshold and the total number of correct assignments to MS^2 (4870) or MS^3 (1256) spectra, respectively. Similarly, the false positive error rate is calculated as the fraction of incorrect matches in the total number of spectra above each probability threshold. Note that there is redundancy between the MS^2 and MS^3 peptide assignments, so summing the total possible number of correct peptide identifications from both MS^2 and MS^3 scans would not reflect the total number of unique identifications.

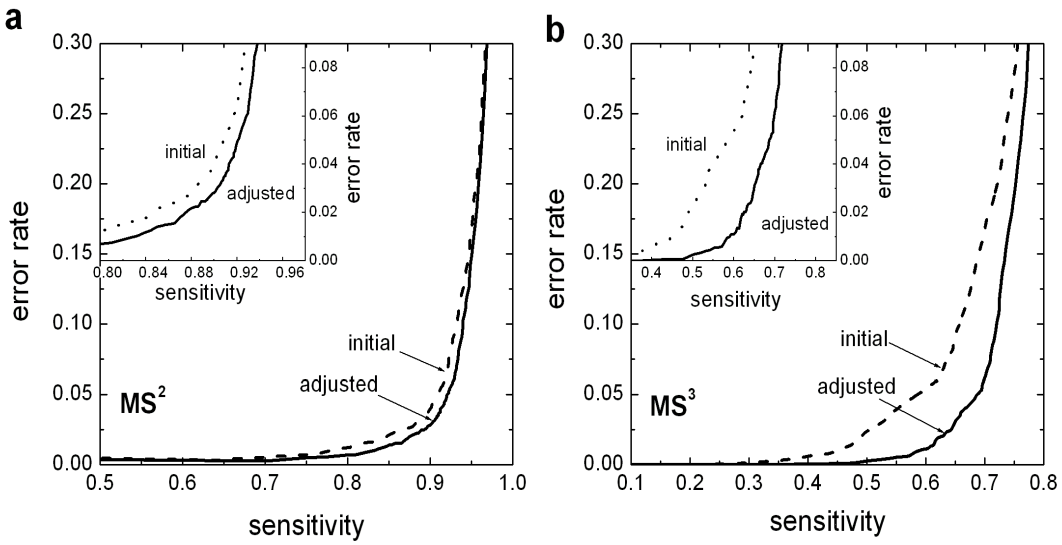


Figure 3-8. Performance of MS^2 and MS^3 scores with probability adjustment. Error rate of MS^2 a) and MS^3 b) scores are shown as a function of sensitivity for initial (dashed) and adjusted (solid) probabilities. Inserted panels are zoomed areas of the plots for the 0 - 10% error rate range

For both the MS^2 and MS^3 scans, the adjusted probability provides a better performance profile, achieving greater sensitivity at an equivalent error rate as compared to the initial data. For example, at a 0.9 probability threshold, the initial MS^2 probability results in the selection of 4072 correct peptide assignments at the expense of 67 incorrect ones. Using the adjusted probabilities, selecting the same number of correct identifications results in only 38 incorrect peptide assignments. The improvement in MS^3 discrimination is even more pronounced, especially in the optimal region of the curve. Using initial probabilities, 1350 correct and 19 incorrect assignments to MS^3 spectra pass the 0.9 threshold. Using the adjusted probabilities, it becomes possible to select the same number of correct peptide assignments with the inclusion of only one false positive.

3.4.5 Combining MS^2 and MS^3 probabilities

The result of the probability adjustment procedure described above is now two adjusted probabilities for each unique linked pair of scans, one each for MS^2 and MS^3. Possibilities for best utilizing both of these scores in selection of correct and incorrect

identifications are now explored. Ideally, a combined scoring approach would provide a greater discriminatory power for selecting correct and incorrect identifications than a subsequent counting of unique matches based on MS^2 and MS^3 taken individually. Two possibilities for utilizing both scores are examined:

$$P_{comb} = 1 - (1 - p_{MS^2})(1 - p_{MS^3}) \quad (4a)$$

$$P_{\max} = \max\{p_{MS^2}, p_{MS^3}\} \quad (4b)$$

where p_{MS2} and p_{MS3} are the adjusted probabilities for the MS^2 and MS^3 scans, respectively, for the same linked pair. The first option is appropriate when the two probabilities can be considered independent, and has been utilized (in a different context, i.e., for combining the evidence from different peptides) for the protein identification problem (33, 38). P_{comb} reflects the probability that at least one of the two peptide assignments, either to the MS^2 or to the MS^3 spectrum, is correct. However, it is obvious that MS^2 and MS^3 spectra, and therefore the probability scores p_{MS2} and p_{MS3} of those spectra, are not fully independent measurements of a peptide in that identical ions will be measured in both spectra. An alternative approach is to select the assignment with the highest probability, P_{max}, thus reducing the likelihood of possible overestimation of the final probability. P_{max} has been used in other similar situations, e.g. in selecting amongst several alternative equivalent peptides (assignments of the same peptide to multiple MS/MS spectra) in the ProteinProphet protein probability score (33), and in Mascot protein-level scoring (24).

Figures 3-9a and 3-9b show the results of counting the number of correct peptide assignments above specified probability thresholds, utilizing all possible scores calculated for a linked pair as the discriminating measure: initial MS^2, initial MS^3, adjusted MS^2, adjusted MS^3, P_{max}, and P_{comb}. Displayed are the results on the set of all unique linked pairs. A comparison of the initial and adjusted probability results for MS^2 and MS^3 again demonstrates an increase in the number of selectable correct peptide assignments at any probability threshold as a result of the probability adjustment. Both P_{max} and P_{comb} scores perform similarly, and provide improved discrimination as compared to the individual measures. Obviously, the primary reason for the performance increase is the fact that the combined score permits the possibility of selecting either the

MS^2 or the MS^3 for any linked pair, thus permitting a pair to be selected as correct if either probability is above threshold. At the 99% probability threshold, for example, the adjusted MS^2, adjusted MS^3, P_{max} and P_{comb} probabilities correspond to 3141, 1050, 3775, and 3807 correct peptide identifications, respectively. Figure 3-9c provides a measure of the rate of false-positives on these data for the most interesting thresholds. The same performance trends are evident: including roughly 40 false positives, specifically 40, 41, 39, and 39 for adjusted MS^2, adjusted MS^3, P_{max}, and P_{comb} measures, respectively, results in selection of 1806, 4139, 4594, and 4762 correct identifications. In all, P_{comb} provides the most discriminative measure.

In addition to analyzing the discriminative power of computed probabilities, one must also assess their accuracy. Probability accuracy plots for the adjusted and combined measures are shown in Fig 3-9d. The adjusted probability scores still provide an accurate representation of true probabilities and fit the 45° line well. The P_{comb} and P_{max} measures perform similarly well. Interestingly, P_{comb} does not overestimate probabilities as one might expect given the dependence of MS^2 and MS^3 level spectra on this data set. Additional analysis would be necessary to determine if this is a general characteristic.

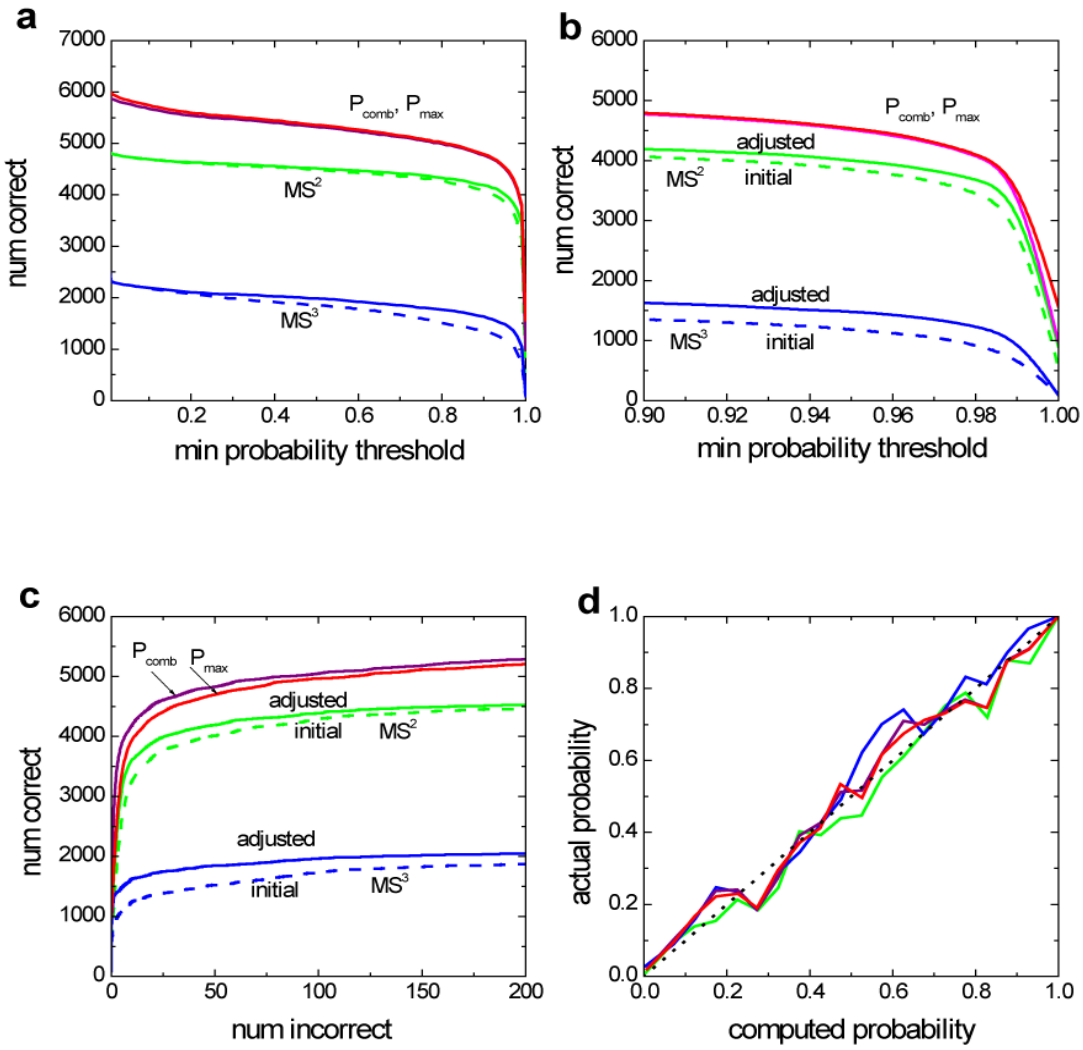


Figure 3-9. **Discriminating power and accuracy of computed probabilities**. a) Total number of correct peptide assignments is plotted as a function of minimum probability threshold for MS^2 and MS^3 spectra alone, both initial and adjusted, and both P_{max} and P_{comb} scores. b) Same as a), zoomed in the region of minimum probability threshold 0.9 to 1.0. c) Number of correct peptide assignments as a function of the number of incorrect assignments, plotted separately for MS^2 (green) and MS^3 (blue) initial (dashed) and adjusted (solid) probabilities, as well as the combined P_{max} (red) and P_{comb} (purple). d) Probability accuracy of the adjusted MS^2, MS^3, P_{max} and P_{comb} probabilities.

3.4.6 Phosphopeptide data set results

One of the main motivating factors in collecting MS^2/MS^3 data is to increase the confidence levels and the total number of phosphopeptide identifications. The identification of phosphopeptides from MS^2 spectra is challenging because spectra

recorded using an ion trap mass spectrometer often exhibit one or more dominant neutral loss peaks of 98 Da, whereas the occurrence and intensity of the other fragment ions (containing peptide sequence information) may be impaired. To investigate potential improvement in discrimination as a result of the probability adjustment on a phosphopeptide-enriched data set, a data set of MS spectra from a single LTQ injection of an IMAC-enriched *D. melanogaster* sample was selected for detailed analysis in this work. The data were acquired in a data-dependent mode, with MS^3 scans triggered for the most abundant peak of the MS^2 spectra which in the case of this sample mostly corresponds to the neutral loss peaks: -98.00 (-116.00), -49.00 (-58.00), -32.60 (-36.66) Da from the precursor, as explained in the Experimental Procedures section. Since the sample in this case is a complex protein mixture, a precise labeling of peptide identifications as 'correct' or 'incorrect' is not possible. Instead, only the composite false discovery rates (FDR) (a single measure for each filtering threshold) can be estimated by counting the number of matches to reversed sequences.

The methodology for generating adjusted probability scores for this data set is analogous to the 9-Mix data set. Top-scoring MS^2 and MS^3 SEQUEST peptide assignments are linked based on consecutive scan numbers, and the top-scoring pair for consecutive scans is selected. Note that if MS^3 spectra are triggered based on neutral loss peaks, charge state ambiguity between matching pairs can potentially be reduced. This fact is not exploited in the analysis; rather, I maintain the same procedure for allowing all possible charge pairs in a match. The match pairs are then classified into sequence match categories as described above. The same four sequence match categories are used: 0, no consecutive match; 1, consecutive match but no matching sequence; 2, matching sequences with MS^3 sequence a subset of MS^2 sequence; and 3, matching sequences with MS^3 sequence identical to MS^2. In this data set, there were only two instances of scans that would correspond to the sequence match category 4: matching sequences with MS^2 sequence a subset of MS^3 sequence. Again, this category was eliminated for simplicity. We note that the additional constraints imposed by the data-dependent triggering of these data and the resultant database searching provisions would allow us to generate additional useful sequence match categories, corresponding to whether the site of modification of a match is identical between the two sequences. We observed a number of instances in

these data where the sequences matched but the sites of modification of the match did not, indicating ambiguity in the localization of the modified residues. A larger data set would allow a more rigorous analysis of these types of results (39, 40).

SEQUEST searching of this dataset produced 16647 and 12218 results for the MS^2 and MS^3 data sets, respectively, corresponding to 7547 unique matching pairs of searched results. Of these, 6270 had non-null MS^3 assignments. Counts for the four sequence match categories are shown in Table 3-4. Most significant is the fact that the sequence match category corresponding to neutral loss-only pairs (match category three) is no longer null; rather it is the more abundant category amongst the two representing matching sequences with 313 unique matches.

	0: No Consecutive	**1**:Consecutive (no match)	**2**: MS^3 Seq In MS^2 Seq	**3**:MS^3 Seq = MS^2 Seq
MS^2 *p(Match|+)*	0.086	0.676	0.079	0.159
MS^2 *p(Match|-)*	0.196	0.796	0.004	0.004
MS^3 *p(Match|+)*	0.000	0.424	0.131	0.445
MS^3 *p(Match|-)*	0.000	0.980	0.015	0.005
Unique Matches	**1277**	**5788**	**167**	**313**

Table 3-4. Match probabilities and sequence match category counts for the phosphopeptide-enriched data set

Corresponding posterior probabilities were calculated for the sequence match categories, and then used to calculate the final adjusted probability for each unique pair. These numbers are also shown in Table 3-4. The frequencies of observing a correct or incorrect assignment to an MS^2 scan with no matching MS^3 sequence (match category one) are relatively close; only a small probability correction occurs for these instances. MS^3 category one probabilities are penalized, as are MS^2 instances that lack a corresponding MS^3 result. A probability boost is received for pairs in categories two and three, with a greater correction given to the latter.

Although a true sensitivity measure for these data is impossible, it is possible to evaluate the relative performance of the various probability measures by examining the number of reversed database matches. The decoy database method is increasingly being

used as an effective means of estimating false positive rates in database searching when other methods of error rates estimation cannot be readily performed (41, 42). At any given probability threshold, the number of matches to reversed sequences can be calculated and compared to the total number of peptide assignments above that threshold to derive an estimate of the FDR (42). A measure of the performance of the various model probabilities on these data is shown in Figure 3-10.

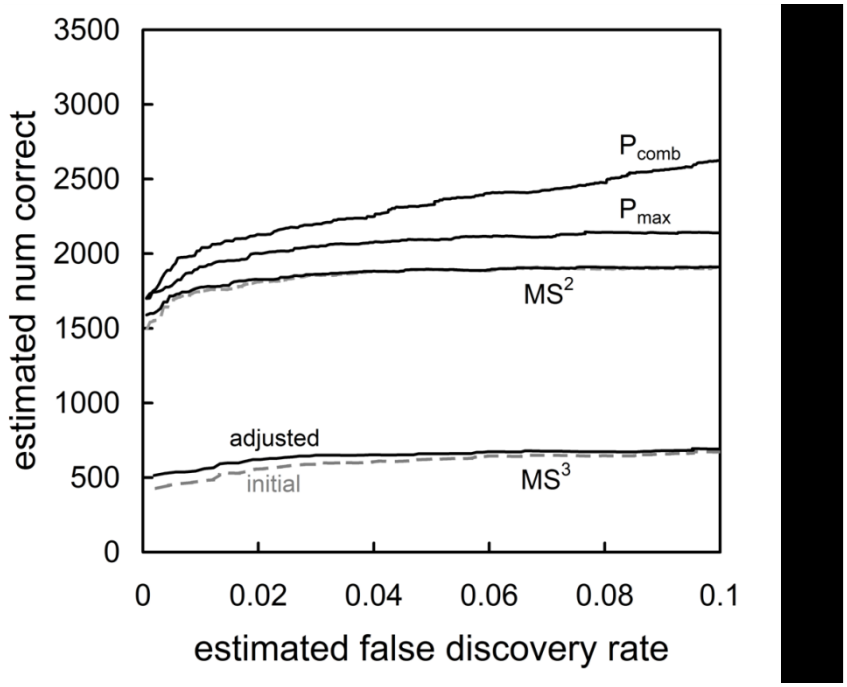


Figure 3-10. **Performance of probability scores on the phosphopeptide data set**. The number of correct identifications estimated using the decoy database method is plotted as a function of FDR estimated using the decoy database search method.

The figure plots the estimated number of correct identifications as a function of FDR. These data are generated by ranking all peptide assignments in order of decreasing probability. The number of assignments of peptides from the forward database (n_f) having a probability equal or greater than the probability of the n^{th} top-ranking reverse entry (n_r) is counted, and the estimated false discovery rate is determined as n_r/n_f. The estimated number of correct assignments is similarly measured as $n_f - n_r$. This analysis is done separately for each of the initial and adjusted probability measures: MS^2 and MS^3 initial and adjusted, as well as the combined probability measures P_{comb} and P_{max}. A

version of these data in table form is provided in Table 3-5, which presents estimated false positive percentages and number of forward match counts for inclusion of one, two, five, ten, fifty, and one hundred reversed matches, as well as the number of those forward entries that are identified as containing phosphorylation sites.

	ms2-initial			ms2-adjusted			ms3-initial			ms3-adjusted			P_{max}			P_{comb}		
# rdb hits	prob cutoff	% false	# forwd (# phos)	prob cutoff	% false	# forwd (# phos)	prob cutoff	% false	# forwd (# phos)	prob cutoff	% false	# forwd (# phos)	prob cutoff	% false	# forwd (# phos)	prob cutoff	% false	# forwd (# phos)
1	0.943	0.13	1500 (1309)	0.908	0.13	1591 (1390)	0.967	0.47	428 (395)	0.927	0.39	515 (468)	0.927	0.12	1704 (1492)	0.931	0.12	1701 (1489)
2	0.921	0.26	1541 (1345)	0.894	0.25	1600 (1398)	0.948	0.89	449 (414)	0.887	0.75	530 (483)	0.908	0.23	1734 (1516)	0.927	0.23	1705 (1493)
5	0.872	0.62	1596 (1396)	0.852	0.61	1638 (1434)	0.868	2.04	484 (444)	0.825	1.81	549 (497)	0.887	0.57	1754 (1536)	0.844	0.55	1800 (1577)
10	0.680	1.16	1715 (1504)	0.644	1.14	1737 (1523)	0.654	3.55	554 (502)	0.523	3.24	607 (543)	0.825	1.09	1817 (1593)	0.636	1.03	1928 (1695)
50	0.293	5.16	1889 (1652)	0.251	5.12	1903 (1662)	0.157	13.4	699 (625)	0.075	12.9	727 (649)	0.319	4.72	2068 (1810)	0.168	4.44	2199 (1930)
100	0.098	9.56	1993 (1746)	0.081	9.55	1995 (1745)	0.070	22.9	774 (690)	0.034	22.7	783 (698)	0.133	8.73	2190 (1924)	0.054	8.08	2375 (2083)

Table 3-5. False positive error rate estimation in the phosphopeptide-enriched data set. For each set of probability scores in the study, the estimated false positive percentages as a function of number of included reversed database matches are listed, as well as the probability corresponding to the nth highest-ranked reversed database match (n = 1, 2, 5, 10, 50, or 100). The actual number of forward database entries that are selected by including the corresponding number of reversed entries are displayed as well. The number of forward entries that are identified as phosphopeptides is listed in parentheses.

As can be seen from Figure 3-10 and Table 3-5, at equivalent false discovery rates, the adjusted probability measures for MS^2 and MS^3 data provide a small but distinguishable improvement in the number of correct entries that can be selected, particularly for MS^3. The bigger benefit of course comes with the combined P_{comb} and P_{max} scores, which provide a much higher selection rate of forward matches than the initial MS^2 and MS^3 probabilities.. For example, by filtering the data using P_{max} instead of the initial MS^2 probability it becomes possible to extract 203 more forward matching identifications without allowing any reverse database matches (1703 peptide identifications vs. 1499). At a roughly 5% FDR, the initial MS^2 probability estimates 1893 correct peptides whereas the P_{max} measure selects 2093. It is interesting that P_{comb} is much more discriminative than the P_{max} probability measure on these data, selecting 2328 correct peptides at the 5% FDR. Overall, the acquisition of MS^3 spectra does appear to increase the total number of phosphopeptide identifications by 10-25% in this data set, depending on the specific combined probability score used for comparison.

The results discussed above for this sample have focused on the total number of identifications, the majority of which are phosphopeptides. An equivalent plot of the results, but including only ranked non-phosphorylated identifications from the phosphopeptide data set, is shown in Figure 3-11. In general, the same trends can be seen; the model improves the assignment scores of unmodified peptides as well.

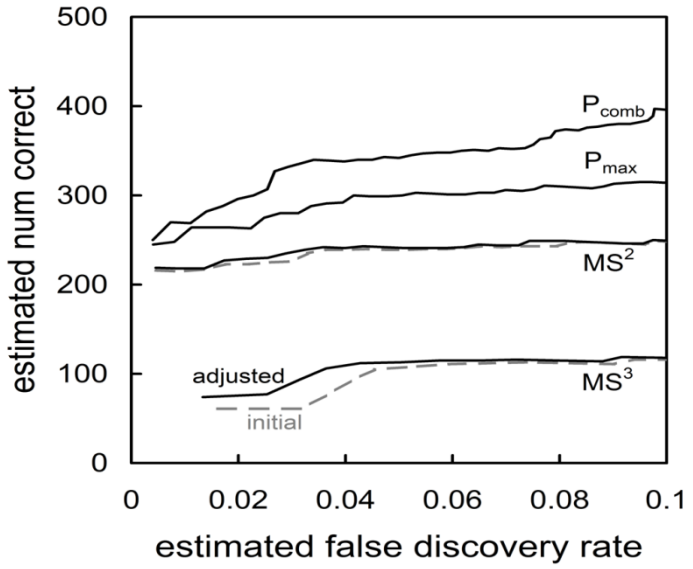


Figure 3-11. **Performance of probability scores on the non-phosphorylated peptides from the phosphopeptide data set.**

3.4.7 Example MS^2 and MS^3 spectra from the phosphopeptide data set

In order to understand the underlying reasons for improved identification confidence, it is informative to briefly revisit the example shown in Figure 3-7. These spectra are representative illustrations of matched MS^2 and MS^3 phospopeptide spectra of various precursor charge states. Several spectral features are of interest. Figure 3-7a shows an example of a +2/+1 match pair. The threonine in position three of the sequence matching the MS^2 spectrum is phosphorylated. The large y_{12} peak corresponding to a fragmentation n-terminal to a double proline was selected by the instrument for MS^3. This is a general characteristic of the singly charge spectra corresponding to correct identifications in these data: the majority are proline-directed, with a Pro identified in the first position. Although the fragmentation is reasonable in this MS^3 spectrum, a large fraction of singly-charged spectra exhibit poor fragmentation with one or two major peaks corresponding to Pro, Asp or occasionally Glu cleavage dominating. This is not surprising due to the relatively low energy imparted to singly-charged ions via collision-induced dissociation (CID) in a trap instrument; typically the most facile fragments are the most readily observable. As can be seen, many of the same ions occur in both spectra. However, the shorter sequence and the absence of the phosphorylated residue in the MS^3

simplify the spectrum and increases confidence in the identification. Figure 3-7b shows a +3/+1 phosphopeptide example. +3/+1 instances are rarer than the +2/+1 (see Supplementary Figure 3), and the same trends occur. The MS^3 spectrum shown is a proline-directed fragmentation event, with Asp-directed fragmentation peaks dominating the spectrum.

Figure 2d is an example of a +2/+2 phosphopeptide ion. The peak selected for MS^3 corresponds to the doubly-charged y_{13} peak with a -98 Da loss of the phosphate moiety. Although many identical ions are identified in both spectra, there is a significant difference in the fragmentation pattern, with several ions observable in MS^3 which are not readily observable in the MS^2.

3.4.8 Data set dependence of probability adjustment

Since the two primary data sets used in this work differ significantly in terms of sample complexity, it is also informative to compare these two data sets with respect to the MS^2/MS^3 matching statistics and the degree to which the initial peptide probabilities are adjusted to account for the sequence match information. The Match parameter distributions *p(Match|+)* and *p(Match|-)* vary between the data sets, reflecting the differences in the sample complexity and data set size. This is illustrated in Figure 3-12, which plots the logarithm of the ratio *p(Match|+)/p(Match|–)* for each match category *k* for both data sets. A ratio greater than 1 (log ratio greater than 0) indicates the region where the probabilities are boosted after adjustment for Match information, whereas a ratio less than 1 (log ratio below 0) indicates that the Match adjustment reduces the probability that a peptide assignment is correct. While the overall trend is similar for both data sets, significant differences exist in the amount of adjustment. For example, the penalty applied to a peptide assignment to a MS^2 spectrum with no subsequent MS^3 spectrum (match category 0) is approximately twice as high in the case of the phosphopeptide enriched data set than in the 9-Mix data set. On the other hand, the amount of probability boost for peptide assignments in the Match=2 category is higher in the case of the 9-Mix data set. A better understanding of these results requires analysis of

the MS^2-MS^3 linking statistics for a larger data set. However, it is clear that the amount of probability adjustment in each sequence match category is data set-dependent. Thus, it is advantageous to use statistical methods for combining MS^2 and MS^3-level data that can learn the appropriate amount of probability adjustment from the data itself, such as the method presented in this work.

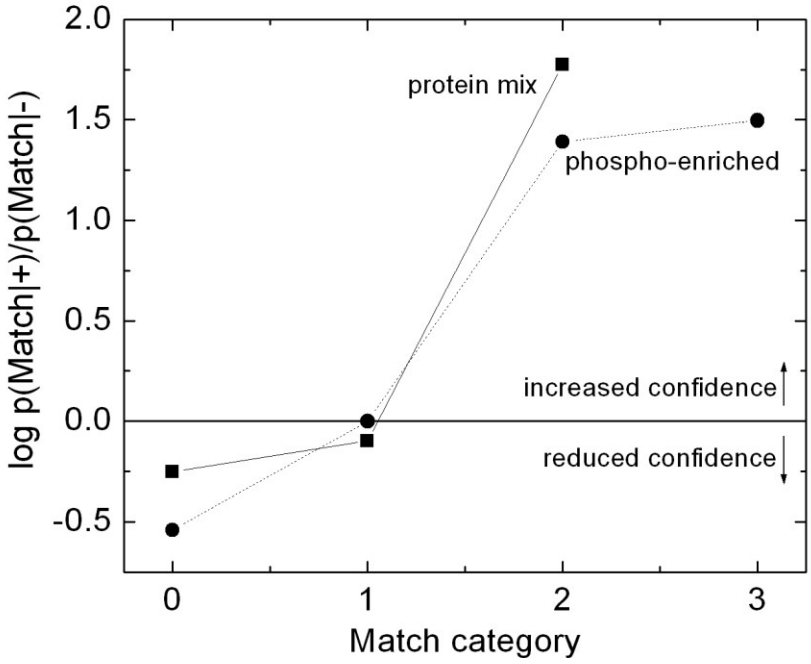


Figure 3-12. **Degree of probability score adjustment by sequence match category for the 9-Mix and phosphopeptide data sets**.

3.4.9 Comments on the overall merit of generating MS^3 data

This chapter describes a method for utilizing MS^2 and MS^3 information for cases in which such data has been generated. A fundamental question arises, however, as to whether or not the benefits of generating MS^3 justifies the additional cycle time on the instrument, or whether the additional MS^2 spectra that would be generating in that time would offset the potential advantage. It has recently been suggested (e.g. Ref 43) that the overall benefit of generating MS^3 information for phopsphopeptide experiments may be limited. Although a comprehensive analysis of the merits of MS3 data generation is

beyond the scope of this work, the situation is explored here by comparing sets of mass spectrometry runs on identical samples utilizing both methods: the MS^2/MS^3 cycle discussed above, and an MS^2-only method.

LC-MS/MS analysis was performed on two additional IMAC-enriched whole-cell D. melanogaster tryptic digests using a Thermo LTQ, as described in Experimental Procedures. Each sample was separated into two equal fractions which were run individually using the MS^2/MS^3 run method or the MS^2-only method. MS^2 and MS^3 peaklists were extracted from the raw data file and searched separately using SEQUEST. Final SEQUEST reports were then combined into two final result sets for each pair of experiments, one set for the MS^2/MS^3, and one for the MS^2-only data. These four result sets were then analyzed using Peptide/ProteinProphet.

To compare results at both the peptide and protein levels, individual identifications for each of the two final result sets were grouped based either on unique peptide sequence or protein accession numbers. The union, intersection and differences between the MS^2/MS^3 and MS^2-only runs were calculated. The results are displayed as Venn diagrams in Figure 3-13 for both pairs of experiments. Given that there was significant variation between the number of peptide and protein identifications of the same run method, the two pairs of experiments were not combined to reduce the effect of instrument sampling rate variability in peptide identification, providing a more fair assessment of differences between the two methods. The top pair of Venn diagrams indicate the number of unique proteins identified by each method. Proteins were included in a set if they participated in an identified protein group (see Ref 33) with a group probability of at least 0.95. Proteins from the same group (indistinguishable proteins given the sequences of identified peptides) were counted as a single entry. The lower set of Venn diagrams shows unique peptide identifications. Peptides were included in these sets if their modified sequences were unique, i.e. two peptides with any modification or sequence differences were considered two unique peptides for the main figure. PeptideProphet probability scores of 0.95 or above were required for inclusion. Peptide uniqueness can be defined by a number of standards, however, and the number of identifications listed in each area of the Venn diagram may be overestimated depending on the definition. The break-out boxes for each of the peptide sets indicate the number

for each region of the Venn diagram under four alternative definitions of peptide uniqueness. Under the Type 1 definition, peptides identified from consecutive MS^2 and MS^3 scans that differ only by the loss of one or more phosphate groups on one of the residues (i.e., MS^3 was triggered on the neutral loss) were considered identical and counted as one. Under the Type 2 definition, peptides which differ at the N- or C-terminus by one or more amino acid residues (e.g., due to a missed cleavage) were considered identical, e.g.

FVS+80EGDGGHVKPTTF
FVS+80EGDGGHVKPTTFTMR
FVS+80EGDGGHVKPTTFTMRD

where S+80 indicates a phosphorylated Ser residue. Under the Type 3 definition, peptides were counted as identical if they had the same sequence but the modification site was ambiguous (residues identified as being phosphorylated are within three amino acid sequences of each other), e.g.

KES+80NSEDELEYDPSLYPQR
KESNS+80EDELEYDPSLYPQR

Under the Type 4 definition, peptides were counted as unique based on the sequence alone, e.g.:

KKES+80NS+80EDELEYDPSLYPQR
KKES+80NSEDELEYDPSLYPQR
KKESNS+80EDELEYDPSLYPQR
KKESNS-18 EDELEYDPSLYPQR

were considered identical sequences. While these four definitions do not include all possible types and permutations that occur, using them to count peptides allows a more comprehensive comparison between the data sets.

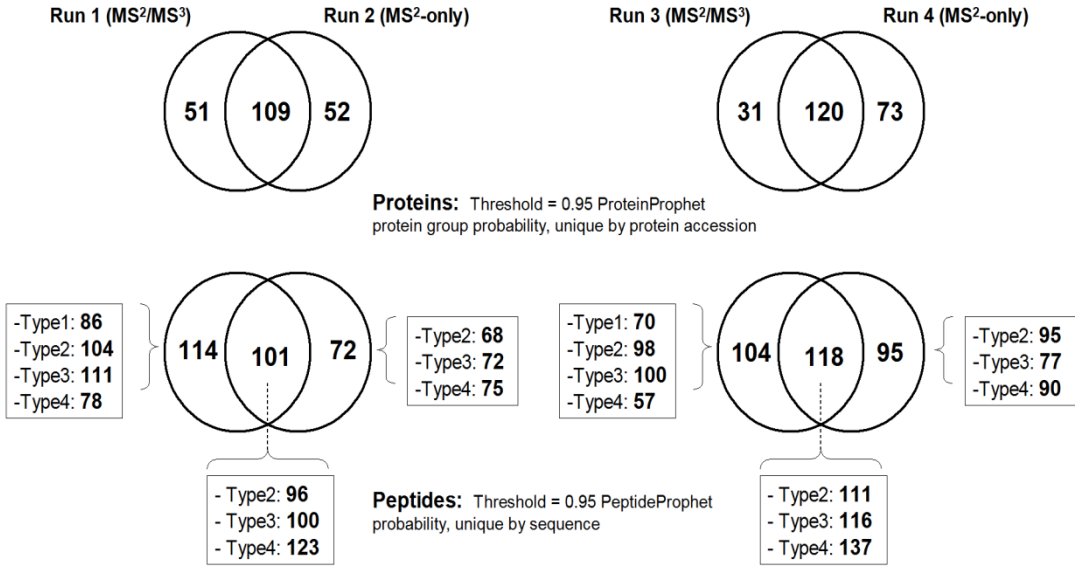


Figure 3-13. **Comparison of MS^2/MS^3 and MS^2-only experimental runs**. Two equivalent pairs of runs are shown, labeled Run1 – Run4. Venn diagrams display overlap between MS^2/MS^3 (left) and MS^2-only (right) data sets based on unique identifications at the peptide and protein levels. All identifications are based on a 95% probability threshold. The top diagrams display protein identifications based on unique Uniprot entry name. The numbers represent the number of ProteinProphet protein groups that have a protein group probabilities equal or greater than 0.95. The lower diagrams show the same for peptide identifications based on peptide sequence, using initial PeptideProphet probability scores. Peptide identifications with the least stringent, most inclusive uniqueness criteria are shown in the main figure. Counts for each region of the diagram utilizing more stringent uniqueness criteria are shown in the boxes, labeled as "- Type".

The results indicate that for these data there are potential advantages to both techniques. At the protein level, the majority of proteins were identified by both methods. However, in one pair of runs the MS^2-only method outperformed the MS^2/MS^3 method by identifying 42 more unique proteins than the MS^2/MS^3 method. At the peptide level, the MS^2/MS^3 method was able to identify more phosphorylated peptide forms in both sets of runs under most of the criteria in which modifications were considered unique (Types 1-3). In terms of the number of unique peptides identified by sequence-alone (Type 4), not taking into account modification state, the MS^2-only set identifies more peptides in one of the runs. This suggests that, at least for certain conditions, sequence coverage may be better with the MS^2-only method.

Overall, these results indicate that generation of MS^3 data may result in a decrease in the number unique peptide and protein identifications. However, several additional comments are necessary for more objective evaluation of the benefits of acquiring MS^3 data. First, the probabilities used in the comparison presented above (Figure 7) were the original probabilities generated by the PeptideProphet and ProteinProphet tools. The probability correction procedure described in this work should permit the selection of a greater number of peptides (and therefore proteins) at a fixed FDR, which would potentially mitigate the loss of sequence coverage. Furthermore if the goal of the study is to identify as many unique modification states as possible, MS^3 data may improve the results. It should also be mentioned that the phosphopeptide data sets used in this work were of high quality (high degree of phosphopeptide enrichment), resulting in sufficiently strong intensity MS signal of phosphopeptide ions and relatively good MS2 fragmentation. On the other hand, it is possible that in other data sets (e.g., no or poor phosphopeptide enrichment), the relatively low abundance of phosphorylated peptides would lead to less intense MS signal and less interpretable MS^2 spectra, thus making benefits of acquiring MS^3 data more apparent.

3.5 Concluding Remarks

The generation of MS^3 information is common in directed areas of proteomics such as phosphopeptide identification. Whether generation of MS^3 information is the best strategy or not is partially dependant on the overall goals of the experiment. Data generated from a complex phosphopeptide-enriched sample suggest that generation of MS^3 spectra can potentially result in an increased number of unique phosphorylation site identifications. On the other hand, the cycle time spent on generation of MS^3 data does appear to detract from the overall number of unique peptides (by sequence only) and proteins identified in such an experiment. Also, although MS^2 spectra in which neutral loss peaks are dominant are still observed in current generation trap instruments, these spectra appear to frequently contain better backbone fragmentation than older equivalents due to increased ion capacity of the trap. Nevertheless, in experiments in which MS^3 data

have been generated, MS^2/MS^3 matching information from the entire experiment can be used to adjust the probabilities of the individual peptide assignments, which has the effect of compensating for the reduced number of MS^2 spectra.

In cases in which a very high certainty in a mapped phosphorylation site is needed, MS^3 experiments are highly valuable as exemplified in the mapping of phosphorylation sites for which biological follow-up experiments are performed. Also, in cases in which neither measurement time nor the amount of phosphopeptide samples are limiting factors, the measurement of MS^3 spectra is advantageous. In fact, in an experimental setup which aims to maximize the number of identified phosphorylation sites from a complex sample, one efficient strategy is to first perform MS^2 experiments and then target specifically the unidentified phosphopeptide ions using MS^2/MS^3 measurements (44, 45).

Generally speaking, much of proteomics data analysis relies on the scores and probabilities produced by automated search algorithms. It is thus important that any probability measure is accurate, and makes use of all available information, particularly in situations where the targeted peptide identifications are rare, e.g. for phosphopeptides and/or when proteins are identified by a reduced number of peptides (such as an analysis in which N-terminal peptides are enriched). Here I have described methods for translating the additional information obtained by matching coupled peptide assignments to MS^2 and MS^3 spectra into a combined probability score, improving the ability to discriminate between true positive and false positive identifications. I have demonstrated an increase in sensitivity and a corresponding decrease in the error rate of selecting correct identifications as a result of the adjusted probability using a mixture of known standard proteins, and applied the method to a complex phosphopeptide-enriched data set, demonstrating an improved discrimination between correct and incorrect peptide assignments for that sample.

The goal of this study was to describe a relatively simple but valid mechanism for adjusting probabilities of peptide identifications in scenarios in which standard database searching has been performed on MS^2/MS^3 data sets. An alternative computational strategy for accommodating MS^3 information is to merge MS^2 and MS^3 spectra into a single spectrum prior to database searching. Full investigation of the relative merits of

pre-database search, spectral merging approaches versus a post-database search probability adjustment procedure such as the one discussed here is beyond the scope of this work, but is the subject of current investigation. Other methodologies, such as merging spectra from differently charged precursors of the same peptide, could likely be utilized to improve peptide identification as well.

As instrumentation continues to improve the speed and accuracy of tandem MS measurements, the ability to generate complementary information such as MS^3 spectra for any given ion will become increasingly practical. Methods for accommodating this information are consequently useful, and can significantly improve the quality of the results generated by automated processing of mass spectrometry data.

3.6 Data and Code Availability

MzXML and RAW datafiles, and processed unique linked pair data, for both the 9-Mix and phospho samples are available online via the Tranche system (http://tranche.proteomecommons.org). The software used in this work was developed in Python. Python modules were implemented making use of the code library available with the InsPecT software package by the UCSD Computational Mass Spectrometry Research Group (28). All code modules developed by myself for this project are available upon request.

References

1. Hager, J. W. (2002) A new linear ion trap mass spectrometer. *Rapid Commun in Mass Spectrom* 16, 512-526

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

3. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem Rev* 101, 269-295

4. Nesvizhskii, A. I. (2006) Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* 367, 87-120

5. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 4, 1419-1440

6. Sadygov, R. G., Cociorva, D., and Yates, J. R., 3rd (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 1, 195-202

7. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5, 699-711

8. Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes J. S., Gruissem, W., Baginsky, S., Aebersold, R. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* 5, 652-70.

9. Pevzner, P. A., Mulyukov, Z., Dancik, V. and Tang, C. L. (2001) Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Res.* 11: 290-299.

10. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., and Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol* 7, R80

11. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 101, 13417-13422

12. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 101, 12130-12135

13. Bodenmiller, B., Mueller, L. N., Pedrioli, P. G. A., Pflieger, D., Jünger, M. A., Eng, J., Aebersold, R., and Tao, W. A. (2007) An integrated chemical, mass

spectrometric and computational strategy for (quantitative) phosphoproteomics: Application to Drosophila melanogaster Kc167 Cells. *Molecular BioSystems*, 3, 275-286.

14. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 4, 310-327

15. Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol Cell Proteomics* 5, 949-958

16. Zabrouskov, V., Senko, M. W., Du, Y., Leduc, R. D., and Kelleher, N. L. (2005) New and automated MSn approaches for top-down identification of modified proteins. *J Am Soc Mass Spectrom* 16, 2027-2038

17. Zhang, Z., and McElvain, J. S. (2000) De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal Chem* 72, 2337-2350

18. Demelbauer, U. M., Zehl, M., Plematl, A., Allmaier, G., and Rizzi, A. (2004) Determination of glycopeptide structures by multistage mass spectrometry with low-energy collision-induced dissociation: comparison of electrospray ionization quadrupole ion trap and matrix-assisted laser desorption/ionization quadrupole ion trap reflectron time-of-flight approaches. *Rapid Commun Mass Spectrom* 18, 1575-1582

19. LeDuc, R. D., Taylor, G. K., Kim, Y. B., Januszyk, T. E. Bynum, L. H., Sola, J. V., Garavelli, J. S., Kelleher, N. L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* 32, W340-W345

20. Frank, A., and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 77, 964-973

21. Goodlett, D. R., Keller, A., Watts, J. D., Newitt, R., Yi, E. C., Purvine, S., Eng, J. K., von Haller, P., Aebersold, R., and Kolker, E. (2001) Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun Mass Spectrom* 15, 1214-1221

22. Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res* 11, 290-299

23. Regnier, F. E., Liu, P. (2002) An isotope coding strategy for proteomics involving both amine adn carboxly group labeling. *J Proteome Res*,1, 443-50

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

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

26. Eng, J. K., McCormack, A. L., and Yates, J. R. r. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976-989

27. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3, 958-964

28. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 77, 4626-4639

29. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2, 1406-1412

30. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., and Aebersold, R. (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nature methods* 4, 231-237

31. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1, 2005.0017

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

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

34. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33:D154-159.

35. Salek, M., Lehmann, W. D. (2003) Neutral loss of amino acid residues from protonated peptides in collision-induced dissociation generates N- or C-terminal sequence ladders. *J Mass Spectrom* 38, 1143-9

36. Martin, D. B., Eng, J. K,, Nesvizhskii, A. I., Gemmill, A., Aebersold, R. (2005) Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal Chem* 77, 4870-82

37. Malmstrom, J., Lee, H., Nesvizhskii, A. I., Shteynberg, D., Mohanty, S., Brunner, E., Ye, M., Weber, G., Eckerskorn, C., and Aebersold, R. (2006) Optimized peptide

separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J Proteome Res* 5, 2241-2249

38. MacCoss, M. J., Wu, C. C., Yates, J. R. 3rd. (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem.* 74, 5593-9

39. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635-648

40. Villen, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* 104, 1488-1493

41. Elias, J. E., Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. Mar 4: 207-14..

42. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2, 43-50

43. Li, X., Gerber, S. A., Rudner, A. D., Beausoleil, S. A., Haas, W., Villen, J., Elias, J. E., Gygi, S. P. (2007) Large-Scale Phosphorylation Analysis of alpha-Factor-Arrested Saccharomyces cerevisiae. *J Proteome Res* 6, 1190-7

44. Picotti, P., Aebersold, R., Domon, B. (2007) The Implications of Proteolytic Background for Shotgun Proteomics. *Mol Cell Proteomics*. May 28 [Epub ahead of print]

45. Domon, B. Aebersold, R. (2006) Mass spectrometry and protein analysis. Science 312, 212-7.

Chapter 4

Optimizing the Phosphopeptide Analysis Pipeline

4.1 Summary

Current mass spectrometers provide a number of alternatives for producing spectra specifically for phosphopeptide analysis. In particular, generation of MS^3 spectra in a data-dependent manner upon detection of the phosphoric acid neutral loss is a popular technique for circumventing the problem of poor phosphopeptide backbone fragmentation. The newer Multistage Activation method provides another option for generating phosphopeptide spectra. Generation of these types of spectra require additional cycle time on the instrument, however, and reduce the number of MS^2 spectra that can be measured in the same amount of time. Additional informatics is often required to make most efficient use of the additional information provided by the spectra produced using these methodologies as well. Here I present a comparison of several mass spectrometry methods commonly used for the study of phosphopeptide-enriched samples: an MS^2-only method, a Multistage Activation method, and an MS^2/MS^3 data-dependent neutral loss method. Several strategies for dealing effectively with the resulting MS^3 data in the latter approach are presented and compared. The overall goal is to infer whether any one methodology performs significantly better than another for identifying phosphopeptides.

4.2 Introduction

Phosphorylation is one of the most highly studied and ubiquitous protein post-translational modifications (PTMs), playing a key role in cell cycle regulation, cell growth and death, metabolism, transcription, morphology and motility, and differentiation due to its prominence in signaling mechanisms and protein complex formation (1, 2). The fact that there are over 500 protein kinases in mammals (3, 4), a number that is likely doubled in plants (5), and that an estimated 30% or more of proteins are phosphorylated at some point during their life cycle (2), underscores the biological importance of this modification. Alterations in normal phosphorylation patterns have been implicated in a number of diseases, including cancer (6-8) and Alzheimer's (9). The identification of phosphoproteins, and understanding the dynamics of this modification in response to cellular and environmental factors, is thus critical for elucidating the systems biology of complex disease mechanisms and global regulatory networks. Consequently, development of methods for detecting and characterizing phosphorylated proteins has been an active area of research in the proteomics community. In particular, high throughput analysis of phosphorylation using directed enrichment methods followed by mass spectrometry has become a standard approach for phosphoprotein detection (10-18).

There are several aspects of phosphoproteomics that make it a challenging endeavor. The primary difficulty is one of stoichiometry: phosphoproteins are often expressed in relatively low amounts in a cell, and relatively few of these proteins exist in a phosphorylated form at any one time. Also, enrichment strategies, while improving, are sub-optimal. Thirdly, phopshopeptides can exhibit poor fragmentation in a mass spectrometer. Lastly, informatics approaches for processing the results of phosphopeptide mass spectrometry data are not yet routine.

The third issue mentioned, the poor fragmentation of phosphopeptides, is due to the fact that the phosphate moiety is often the most labile element on the peptide. In the case of collision-induced dissociation (CID), much of the fragmentation energy used to produce a tandem (MS/MS or MS^2) mass spectrum often is absorbed in the dissociation of the phosphate group. The resulting spectrum is often dominated by one or several peaks corresponding to the neutral loss of phosphoric acid, with little other fragmentation

information useful for identification of the peptide sequence (19). This issue has been addressed using data-dependent MS^3 methodologies for generating mass spectra, typically on ion-trap instruments. Subjecting neutral-loss fragment masses to a further cycle of fragmentation often produces a spectrum with much more useful structural information on the peptide (20). Therefore, phosphopeptides have often been analyzed by automated data-dependent triggering of MS^3 acquisition whenever the neutral loss ion of the appropriate mass is detected in an MS^2 spectrum as a dominant peak (16, 21- 27).

Despite the apparent advantages, researchers are beginning to question the merits of generating MS^3 data for phosphopeptide studies that are conducted on the most current instruments. An argument given is that, due to increased trap capacity, spectra generated on current instruments often contain sufficient fragmentation information in their MS^2 spectra to uniquely identify the peptide. Although the dominant peak is most often still due to the neutral loss of the phosphate, sufficient information is contained in the smaller peaks to derive amino acid sequence information (28). Moreover, it is argued that the cycle time spent on generating MS^3 information detracts from the overall number of MS^2 spectra that can be produced, potentially reducing the number of unique identifications. Addressing this concern, an alternative strategy for fragmentation of phosphopeptides has been proposed and made available on current instrumentation, referred to as Multistage Activation (MSA) or "Pseudo MS^n" (29). With MSA, the neutral loss fragment ions in an MS^2 spectrum are activated without a separate isolation/activation cycle on the instrument for neutral loss fragments; the net result is a composite spectrum containing fragmentation ions from both the original MS^2 fragmentation as well as the fragmentation resulting from activation of the neutral loss ions. It has been demonstrated that the search scores generated by automated search algorithms such as Mascot (30) are improved over their conventional MS^2 equivalents, ideally translating into the ability to select more peptide identifications at any given scoring threshold (29).

A consequence of the MS^3 methodologies discussed above is that the downstream informatics of processing MS^3 spectra requires additional consideration, and is often not equivalent to those used for processing MS^2 spectra alone. The first issue is one of redundancy: MS^2 and MS^3 spectra are ideally derived from the same peptide, and may generate matches to identical peptides using database searching tools. The resulting

matches must be integrated in some manner in final reports. Another issue is that the measured precursor masses associated with MS^3 spectra will not always correspond to the masses of appropriate database peptides calculated using the same rules that are applied in the case of MS^2 spectra. For example, in phosphopeptide analyses variable modifications of -18 Da due to loss of phosphoric acid from S or T residues need to be specified for MS^3 spectra, while the normal +80 Da phosphorylation modification on S, T, and Y are used for MS^2. It is computationally inefficient, and an unnecessary source of false positive identifications, to perform a combined search which permits both the -18 Da loss for MS^2 spectra and the +80 Da addition for MS^3 spectra. Moreover, any MS^3 spectrum generated from a peptide in which amino acids are lost in addition to the phosphate moiety, an event that does occur, requires the searches to be conducted in a semi-tryptic manner, preferably allowing for internal ion masses. This is beyond the capabilities of some of the popular database search platforms.

Such considerations have led several investigators to develop their own algorithms for handling MS^3 scans. Olsen and Mann describe a custom scoring algorithm for MS^3 spectra: their final score for a peptide is the product of the Mascot-generated MS^2 probability and the custom MS^3 score, implemented in a modified version of their MSQUANT software[1] (20). Hoffert et al. developed a framework called PhosphoPIC that processes SEQUEST results of phosphoepeptide-enriched samples to allow more effective filtering and post-search compilation of the data (24). In addition to standard database searching, methods for computing a score associated with the specific site of phosphorylation on the peptide have been published. Beausoleil et al. describe an extension of the score developed by Olsen and Mann that calculates a value indicative of whether the specific site of phosphorylation can be localized to a particular residue (31).

In the previous chapter, in the context of discussing my custom MS^2/MS^3 data processing methodology, I performed an initial assessment of the merits of generating MS^3 spectra by comparing an MS^2-only methodology to an approach generating MS^3 spectra on a phosphopeptide-enriched sample. Initial results indicated that generating MS^3 spectra could indeed result in fewer unique peptide identifications. However, there was evidence that the number of unique sites of phosphorylation increased as a result of MS^3. The goal of this chapter is to address this issue more precisely. Overall, I seek to

provide evidence as to whether a particular methodology for analyzing phosphopeptide data performs better than another, both in terms of the number and quality of identifications as well as in simplicity of implementation. In as equivalent a manner as possible, I compare three approaches: an MS^2-only methodology, an MS^2/MS^3 methodology, and an MSA methodology. The primary criterion for comparing methodologies is the number of unique peptide identifications. I also examine the effect of the method on the determination of the site of modification as well as whether there is an effect on the overall mass accuracy of the results. For MS^2/MS^3 data, I also compare several different informatics methods for interpretation of the resulting spectra.

4.3 Experimental Procedures

4.3.1 Sample Preparation and Mass Spectrometry

The sample used in this study is a trypsin-digested, IMAC-enriched cytosolic protein extract from *Drosophila melanogaster* Kc167 cells, identical to the "Phosphopeptide Sample" described in Chapter 3. The preparation of the phosphopeptide samples is described in detail in Bodenmiller et al. (32). A total peptide amount of ~12 μg was divided into six equal samples, with duplicate samples run on the instrument for each of the three mass spectrometry methods described next.

Mass spectrometry: Mass spectrometry and chromatographic separation of phosphopeptides was basically identical as described in Bodenmiller et al. (32) and Chapter 3, except that an LTQ-FT (ThermoFischer Scientific, Bremen, Germany) and a gradient from 2 % to 25% acetonitrile in 90 minutes was used.

Three mass spectrometry methodologies are explored in this study: a standard MS^2-only methodology (abbreviated as "MS2" in this document), a MultiStage Activation (MSA, also called "Pseudo MS^n") methodology, and a method which generates both MS^2 and data-dependent MS3 spectra (abbreviated "MS3"). For the MS2 methodology, all peptides eluting from the column were recorded in MS mode in the first scan event using the Fourier transform (FT) analyzer. The most intense ion was selected for a product ion spectrum (MS^2) in the second event on the linear trap analyzer of the

instrument. This MS^2 event is repeated up to six times for subsequently less intense peaks exceeding a threshold of 500 ion counts, for a total possible cycle of seven events. For the MSA methodology, all peptides eluting from the column were recorded in MS mode in the first scan event using the FT analyzer. The most intense ion was selected for a product ion spectrum (MSA) in the second event on the linear trap analyzer of the instrument. During fragmentation, the neutral loss species at 98, 49, 32.66 and 24.5 m/z below the precursor ion mass were additionally activated as described in (29). This MSA event is repeated up to five times for subsequently less intense peaks exceeding a threshold of 500 ion counts, for a total possible cycle of six events. Finally, for the MS^2/MS^3 method, first and second method events were analogous to the MS2 method. However, in the third event an MS^3 spectrum was triggered specifically in the event of a phosphate neutral loss (-98 Da for singly, -49 Da for doubly and -32.66 Da for triply and -24.5 Da for quadruply charged peptides). The second and third events are then repeated two more times in the cycle, for the second and third most abundant MS^1 ions, for a total cycle of seven events. A threshold of 200 ion counts was used for triggering an MS^2 attempt. Wideband activation was enabled for all MS^2 and MS^3 scan events. MS^2 isolation width was set to 2 m/z and MS^3 isolation width was set to 4 m/z. For triggering an MS^3 event the most intense ion had to be above 50 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

A profile of spectral counts for each of the six instrument runs is shown in Table 4.1. Charge state could be determined to a high degree of accuracy due to the high mass accuracy of the FT. Singly-charged peaklists were excluded from further analysis in the extraction phase because of their high false positive likelihood and are not shown in the figure. Note that spectra with charge states greater than eight are not considered by Mascot.

Filename	Total Spectra	MS2 Spectra	MS3 Spectra	+2	+3	+4	+5	+6	+7 more
B07-10186_c_MS2	4274	4274	0	1837	1790	476	115	36	20
B07-10187_c_MS2	4321	4321	0	1875	1826	464	109	32	15
B07-10189_c_MSA	3925	N.A.	N.A.	1693	1662	414	99	29	27
B07-10190_c_MSA	3941	N.A.	N.A.	1712	1659	424	102	31	13
B07-10192_c_MS3	5531	3506	2025	1487/951	1498/956	373/97	109/15	24/5	14/1
B07-10193_c_MS3	5554	3507	2047	1449/928	1518/1006	394/87	97/19	31/6	18/1

Table 4-1. Spectra counts for the MS2, MSA, and MS3 datasets. The final term in the filename specifies the type of run: MS2 for MS^2-only, MSA for Multistage Activation, and MS3 for MS^2/MS^3 runs. Columns indicating the number of spectra of each precursor charge state are indicated; designations such as 373/15 represent the number of MS^2/MS^3 spectra of that charge state. Singly-charged peaklists were eliminated from the analysis.

4.3.2 Database Searching and Results Analysis

MzXML files were generated from binary ThermoFinnigan *.raw files using the ReAdW tool available in the Trans-Proteomic Pipeline (TPP) suite of programs (33-35). Peaklist files in *.dta format were extracted from the mzXML files using mzXML2Other tool: for MS^2/MS^3 runs, the -level option[12] was used to extract MS^2 and MS^3 peaklists separately. The database for the phosphopeptide-enriched samples consisted of all *D. melanogaster* sequences exported from the UniProt database (36), 26311 entries total, to which the reversed set of sequences was appended. Results were searched using Mascot with parameters for the MS^2, MSA and summed MS^2/MS^3 (described below) spectra as follows: peptide tolerance of 25ppm, fragment ion tolerance of 0.8 Da; full trypsin digestion, two possible missed cleavages; fixed carbamidomethyl modification of 57.02 for Cysteine; variable modifications of +80 Da for Ser, Thr, and Tyr, and +16 Da for Met. The instrument type was set to ESI-TRAP, this is significant in that it determines the types of ions that Mascot scores. The ESI-TRAP setting does not calculate internal

[1] http://tools.proteomecenter.org/software.php

ions. A comparison of the performance of this instrument setting with a setting that utilizes internal ions showed that the latter significantly degraded the quality of search results (data not shown). The MS^3 spectra were searched with the same set of parameters except that variable modifications of -18 Da on Ser and Thr (instead of +80 Da) were specified to accommodate loss of phosphoric acid leading to a dehydroalanine or dehydrobuyric acid, respectively.

Search results generated by Mascot in the *.dat format were converted to pepxml format using the TPP mascot2xml program. The results were then analyzed using PeptideProphet (34), resulting in a probability score for each spectrum. PeptideProphet was run with the "-l" option, which results in alternate processing of DeltaCn scores marked with '*', representing cases in which the top and second-highest ranked peptide assignment to a spectrum have homologous sequences (>70% sequence identity) as often occurs for phophopeptides when the specific site of phosphorylation is ambiguous. ProteinProphet (35) was also run on the pepxml result files from PeptideProphet for each of the MS2, MSA, and MS3 cases, combining results of the two replicate experiments for each method type for a total of three final result sets at the protein level.

4.3.3 MS^2/MS^3 spectra data processing

MSA spectra are composite spectra containing MS^2 fragment ions as well as ions generated by activating several neutral loss product ions generated by the initial MS^2 event. These spectra require no special processing to search using standard tools. However, it has been noted (29) that current search engines may not consider all combinations of fragment ions that may be simultaneously present in MSA spectra (e.g. simultaneously scoring a corresponding b_n and a b_n-H_3PO_4 ion).

Spectra generated in an MS^2/MS^3 methodology can be processed in several ways. The work presented in Chapter 3 explores a method in which consecutive MS^2 and MS^3 spectra can be paired after database searching, providing an ability to adjust the probability scores of both spectra based on whether the sequence assignments of these consecutive matching pairs of spectra match. The adjusted probability scores for the

paired MS^2 and MS^3 spectra may then be used individually, or the two scores may be combined to form a single score for each unique matching pair. This method performs optimally when MS^2 and MS^3 spectra are searched separately, to avoid the need for specifying multiple variable modifications (+80 Da for MS^2 and -18 Da for MS^3) on the same residue. The specification of the -18 Da neutral loss for MS^3 is required to allow a standard search engine to select peptides of the correct precursor mass in a database. This requirement can be eliminated, however, by simply replacing the precursor mass of MS^3 spectra with the precursor mass of their parent MS^2 scan prior to searching. Although the replaced MS^3 precursor mass in a peaklist or MzXML file no longer reflects the actual mass selected for fragmentation in the MS^3 spectrum, it eliminates the need to search MS^2 and MS^3 spectra separately. This "trick" is implemented by default in current BioWorks software distributed with a ThermoFinnigan instrument (and, to this author's knowledge, is undocumented).

There is yet another method for utilizing MS^3 spectra that can be performed prior to database searching. This method involves combining consecutive MS^2 and MS^3 spectra into a single spectrum. The advantage of this technique is that it has the potential to both increase the signal-to-noise ratio of common ions in both spectra as well as provide more extensive fragmentation information in a single spectrum. Such spectra are referred to as "summation spectra", or just "sum spectra" in the remainder of this manuscript. For this work, sum spectra are assembled by first extracting MS^2 and MS^3 peaklists separately from a raw datafile into separate folders. The scan numbers of the peaklists are then examined to identify consecutive MS^2/MS^3 pairs, as described in Chapter 3 and (25). Consecutive spectra are then merged as follows: First, the intensities of peaks in the two spectra are normalized by comparing the base peak intensity of each spectrum, adjusting the MS^3 peak intensities to match MS^2 levels. Then a new peaklist is constructed composed of peaks from the individual spectra. If two peaks from the MS^2 and the MS^3 spectra are within a specified tolerance (0.4 Da for all data presented here), the peaks are combined by adding the intensities. A single m/z value—the value from the most intense of the two peaks— is used in the sum spectrum for the peak.

A note on terminology: for the remainder of this manuscript, the terms MS^2 and MS^3 will be used when referring generally to spectra of these types, and MS2 and MS3 when referring to specific datasets discussed in this thesis.

4.3.4 Calculating a score for phosphopeptide site localization: a custom Ascore

In a recent publication (31), Beausoleil et. al. describe an algorithm that computes a probability of phosphorylation being localized to a specific site on a peptide, called an Ascore[13]. A custom version of this algorithm was implemented in the Python programming language for this project. There are several differences between my implementation and the published version. Most importantly for this work, my implementation allows me to calculate a score for MS^3 spectra, a feature not supported in the online version of the Ascore algorithm provided by the Gygi lab. Also, the published implementation calculates multiple scores for multiply-phosphorylated peptides. My algorithm functions by searching through all possible permutations of phosphorylated forms given the identified number of phosphates modifying the peptide. It then selects the two highest scoring forms to use in a more refined localization score calculation, generating a single score for the peptide indicating the likelihood that the given modified peptide is the correct form. Thirdly, my version of the algorithm produces slightly different scores than the published version, likely due to differences in the assignment of labels to ions in the spectrum. Given that the use of my custom implementation is for comparative purposes, specifically to determine if the MS methodologies differ in their ability to localize sites of modification, I believe it is a valid and useful representation of the algorithm. However, in that it is not an identical measure, the score produced by my algorithm is called a 'localization score' rather than an 'Ascore' in this document.

[13] http://ascore.med.harvard.edu/

4.4 Results and Discussion

4.4.1 Comparison of the number of identifications

A general belief regarding the merits of generating MS^3 data is that doing so reduces the number of unique peptide identifications, simply by lowering the number of MS^2 scan events in a run. This appears to be the case in some situations, and was confirmed in earlier data presented in the previous chapter. For the more extensive datasets examined in this study, however, this result was not found to be the case. Figure 4-1 shows the number of unique peptide identifications for each of the six instrument runs. Unique matches here are defined by primary peptide sequence; no accounting of modification state of the peptide assignments is taken into account. Peptides were selected at a 0.95 probability threshold as reported by PeptideProphet; a similar number of identifications were obtained by utilizing the Mascot significance threshold for each peptide of (ionscore – identityscore) >= 0.0 (data not shown). Overall, the method performance between all six runs was comparable. However, the MSA methodologies produced more hits than the other two methodologies, 8% more than the MS2 methodologies and 4% more than the MS3.

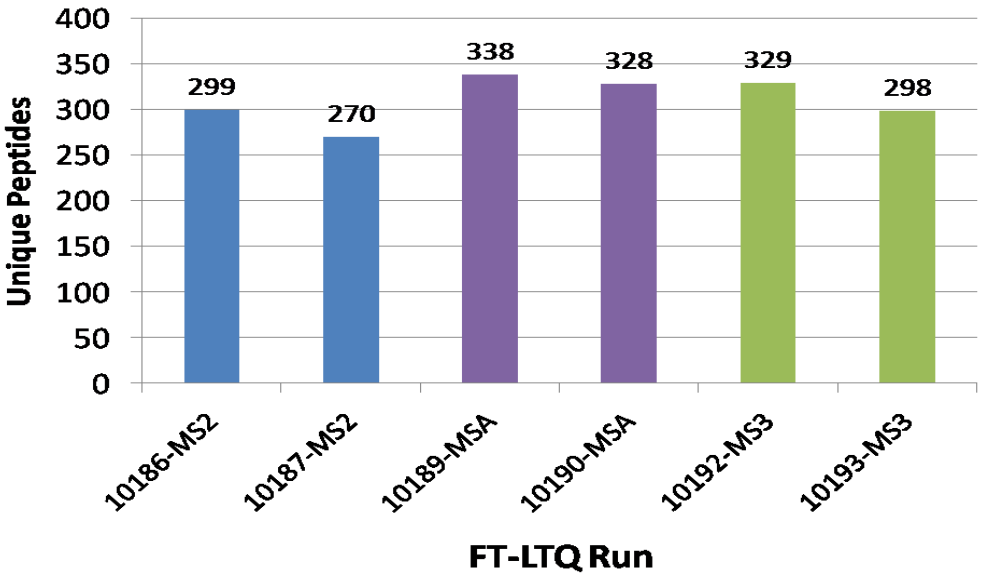


Figure 4-1. Number of unique peptide identified in each dataset. The run method is indicated in the dataset name: MS2-only (MS2), MSA, or the MS^2/MS^3 (MS3) methodology. Peptides selected at a 0.95% confidence threshold as reported by PeptideProphet.

4.4.2 Optimizing the data for MS^3 spectra

The identifications listed in Figure 4-1 for the two MS3 run methods were calculated based on raw, 'unadjusted' probabilities. As described in Experimental Procedures (section 4.3.3), several computational methods exist for further refining the identifications from MS^3 spectra to make efficient use of the information. Results on reported on three different techniques for processing the MS3 datasets: 1. Adjusting the MS^2 and MS^3 probability scores based on the methods discussed in (25) for all identifications in the dataset (notated "MS3-ADJ"); 2. Combining the adjusted probability scores for consecutive matching pairs into a single score using the heuristic: $1-(1-P_{MS2})(1-P_{MS3})$, where P_{MS2} and P_{MS3} are the probabilities of the MS^2 and MS^3 spectra, respectively (notated "MS3-1-p"); and 3. Creating a summed MS^2/MS^3 spectrum for consecutive matching pairs ("MS3-sum").

Figure 4-2 shows the comparison of number of unique peptide identifications produced for the two MS3 runs using each of these methodologies with the raw MS3

scores. The two replicate runs 10192 and 10193 are shown in the same color for each method. The spectra generated for the 10193-MS3 run on the instrument obviously yield fewer identifications. Overall, the three refinement methods produce a modest performance increase over the raw equivalents, resulting in significantly more identifications. The MS3-sum spectra result in roughly the same number of unique identifications as adjusting the MS^2 and MS^3 peptide probabilities individually, suggesting that summing MS^2 and MS^3 spectra can provide a simple method for utilizing MS^3 spectra. Although the MS3-1-p method requires an additional computation step, it is the most successful MS3 technique for identifying unique peptide identifications on these data.

An expanded summary of counts of peptides identified by all of the methods is shown in Table 4-2, including counts for the number of phosphorylated peptides identified. The total number of spectra generated by each method is shown as well as the number of total and unique peptide identifications identified at a PeptideProphet probability threshold of 0.95. A new statistic is calculated, the ratio of Unique peptides identified to Total spectra generated: a U/T or "Utility" ratio. This ratio can be considered a simple measure of spectra quality. It is interesting that the MSA spectra produce the highest scores for this statistic, a reflection of the increase in information resulting from a combination of activation events. An increase is not evident in the summation spectra, however, pointing out a difference in these two techniques.

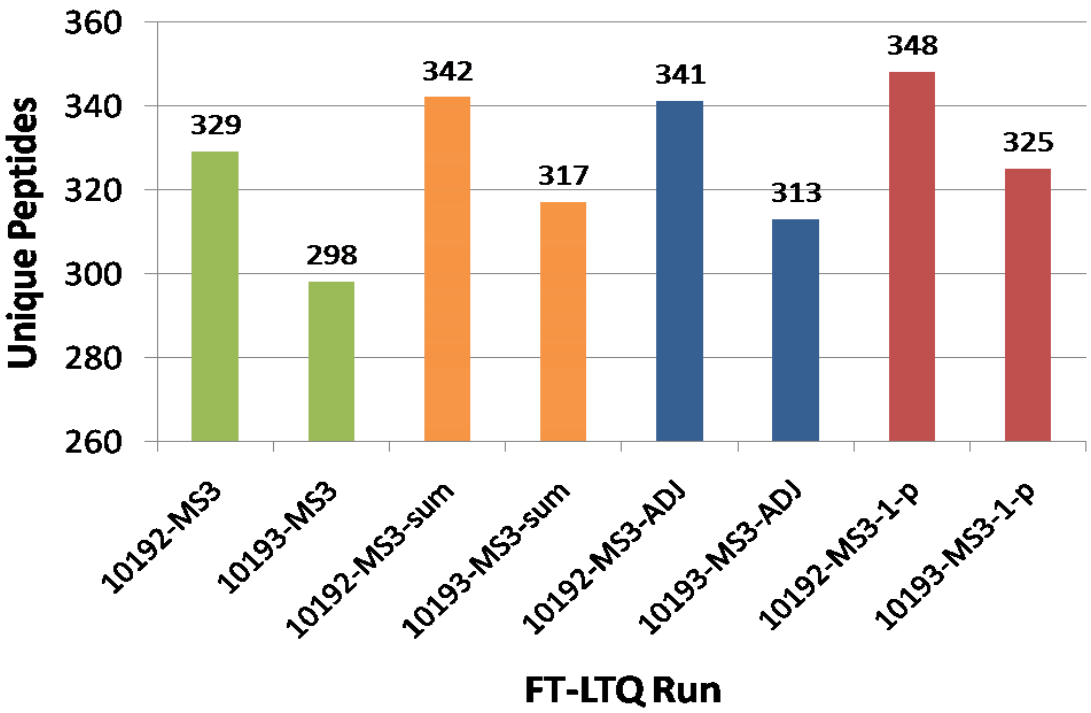


Figure 4-2: **Comparison of MS3 data processing methodologies**. Data are shown for the two replicate runs for which MS^3 spectra were generated (run IDs 10192 and 10193). Results from three alternative refinements, summation spectra (labeled "-sum"), probability adjustment ("-ADJ") and the combined $1-(1-P_{MS2})(1-P_{MS3})$ score ("-1-p") are compared with the unprocessed results.

The number of unique peptides that the search engine identifies as phosphorylated is shown. I also calculate a localization score for every phosphopeptide. I use the score of 19 as the threshold above which a phosphopeptide can be considered localized with high confidence, a value suggested in the original AScore publication. To be considered as a truly uniquely identified form for the Unique Peptides column of Table 4-2, a phosphopeptide must have a localization score above this threshold; otherwise, only the primary amino acid sequence for the peptide is considered. However, the modified sequences of phosphopeptides meeting these criteria are considered unique for the purposes of counting.

Run ID	Total Spectra	Total Peptides	Unique Peptides	Utility	Phospho Peptides	Localized Phosphos
10186-MS2	4274	327	301	0.0704	284	117
10187-MS2	4321	309	272	0.0629	250	106
10189-MSA	3925	385	348	0.0887	305	147
10189-MSA	3941	376	334	0.0848	295	144
10192-MS3	5531	448	338	0.0611	299	129
10193-MS3	5554	413	298	0.0537	277	111
10192-MS3-sum	5531	377	349	0.0631	302	137
10193-MS3-sum	5554	355	321	0.0578	294	127
10192-MS3-ADJ	5531	503	351	0.0635	316	129
10193-MS3-ADJ	5554	472	317	0.0571	302	117
10192-MS3-1-p	5531	386	353	0.0638	311	127
10193-MS3-1-p	5554	364	330	0.0594	304	118

Table 4-2. Summary of peptide identification counts for all methods. The total number of spectra generated for each experiment are shown, as well as total and unique peptides identified at >= 0.95 PeptideProphet probability. The Utility measure provides the ratio of unique peptide identifications to total spectra. The Phospho Peptides column lists the number of unique peptides above threshold that are phosphopeptides. Localized Phosphos list the number of the phosphopeptides identified as having an Ascore >= 19.0.

The ratio of phosphopeptides to total unique peptides in Table 4-2 indicates a high degree of enrichment, confirming the results for the method described in (32). Surprisingly, the MS2-only methodology produces the fewest number of unique peptide identifications. The MS3-1-p methodology and the MSA methodology produce the highest number of unique identifications, 10% more than the MS2 method. The number of localized sites of modification is elevated in the MSA results, even though the MS3-1-p results produce an equivalent number of unique peptide identifications and even more total unique phosphopeptide identifications than MSA. The result is not terribly surprising given that an MSA spectrum can theoretically contain fragment ions from multiple neutral loss ions whereas the MS3 methods contain fragmentation data from at most a pair of MS^2 and MS^3 events. The results suggest a net increase in overall spectral quality as a result, confirmed by the Utility measure. Amongst the four MS3 methodologies shown, although MS3-1-p produces the higher number of unique peptide

identifications and the most overall phosphopeptide identifications, MS3-sum spectra produced the most number of confidently localized phosphopeptide identifications.

4.4.3 Unique protein and peptide identifications for combined datasets

Comparisons were made between the methods' performance in identifying proteins as well. One of the difficulties associated with enriched phosphopeptide samples is that, due to the simplified nature of the mixture, often only one peptide is identified per protein (so-called "one hit wonders"). As a result, a modified selection criteria was used for allowing peptide identifications in the data set. The ProteinProphet tool clusters protein identifications into protein groups based on the underlying peptide identifications, generating a probability for each protein group. All identified groups for a given methodology were ranked by probability score. Since data were searched against a reverse-appended database, a false discovery rate (FDR) could then be calculated for any probability threshold (37, 38). All protein identifications meeting or exceeding the 0.05 FDR probability threshold were thus selected. Based on the information obtained in this clustering, ProteinProphet then produces an adjusted probability score for peptides associated with these protein groups, called the NSPAdjusted probability, which reflects the increase in confidence in peptides that have other "siblings" contributing to a protein identification (35).

Figure 4-3 displays Venn diagrams outlining the number of protein and peptide identifications obtained by the various methods, and their relationships. For these figures, the results from the two replicate runs for each method are combined into a single result. Peptides were included in these figures if their NSPAdjusted probability exceeded 0.50, the decrease in confidence from the 0.95 threshold utilized earlier justified by the fact that the proteins to which they are assigned are known to be correct with high confidence. Only peptides participating in the assignment of proteins which exceed the 0.05 FDR threshold are included. Again, modified forms of phosphospeptides were considered as unique if their localization scores passed the 19.0 significance threshold.

The figures show a large degree of overlap in identifications between the three methods, with greater than half of the identifications for any single method corresponding

with both other methods. There are, however, a significant number of identifications identified by one method alone at the peptide, and thus the protein, level. This difference can largely be attributed to sampling rate on the instrument. The MS2 and MSA methods produce a total of 478 and 519 unique protein identifications, respectively, which include 638 and 716 unique peptides. Of the MS3 methods, MS3-sum outperformed the MS3-1-p method in terms of the number of total identifications.

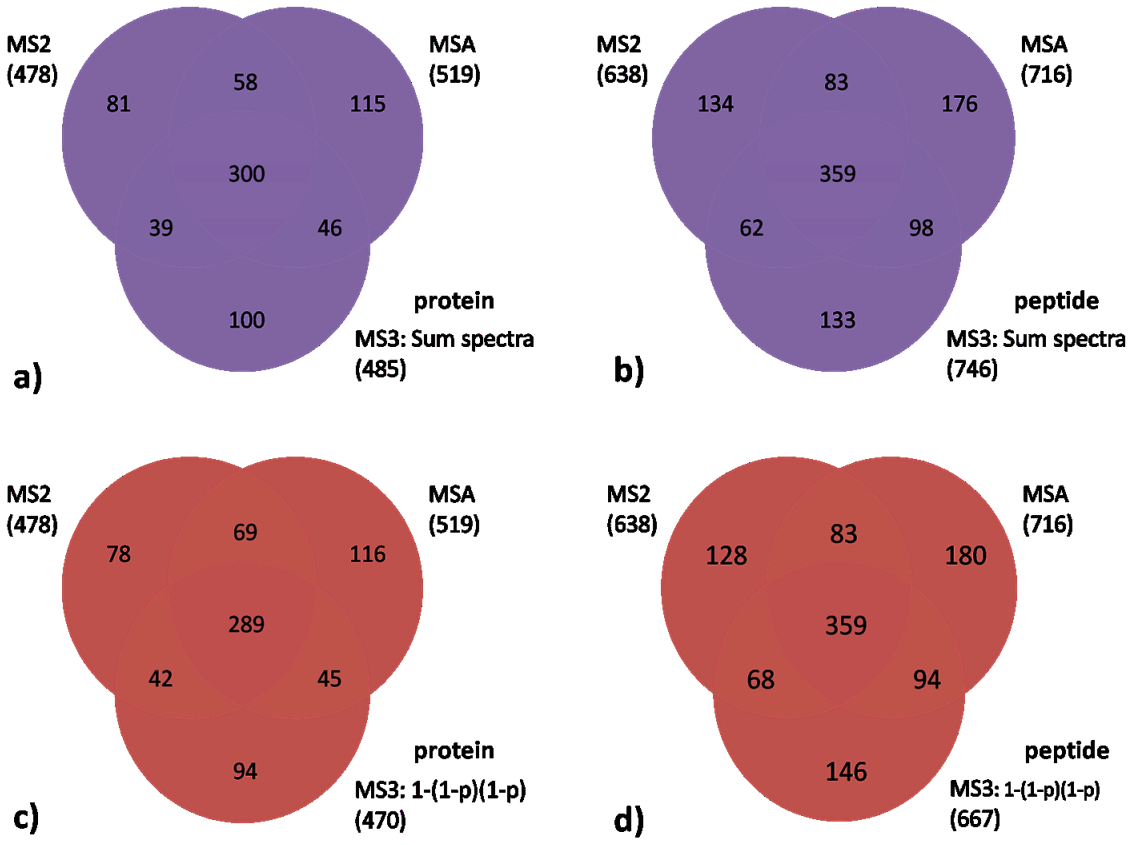


Figure 4-3: **Comparison of protein and peptide identifications for combined datasets**. Venn diagrams of the three primary run methodologies are compared using two MS3 processing methods: MS3-sum spectra (Panels **a** and **b**, top) and the MS3-1-p method (Panels **c** and **d**, bottom). The figures compare unique protein assignments (left) and peptide assignments (right). Total counts for all areas corresponding to each dataset are shown in parentheses. Proteins are selected based on an estimated FDR of 0.05 using the decoy database strategy. Peptides are included for all significant proteins if their NSPadjusted probability scores are >= 0.5. Venn diagrams are not area-proportional.

4.4.4 Effect of methodology on phosphorylation site localization

In addition to the total number of identifications, the methodologies are compared to see if there is a difference in their ability to localize a site of modification. Such a difference might be reflected in a shift in the localization score between the methods. To investigate this, binned localization score values were plotted for the combined results of each individual method. The results are plotted in figure 4-4.

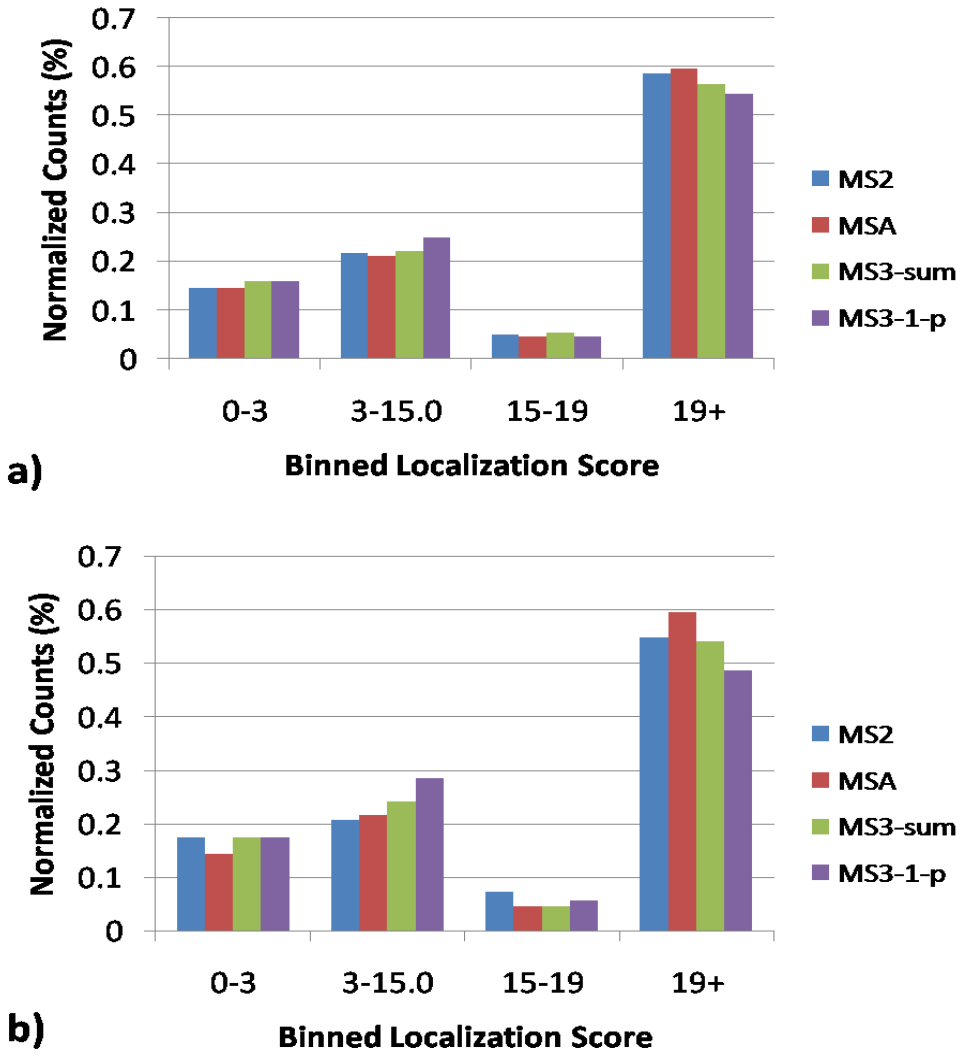


Figure 4-4. Localization score histograms for individual run methods. Each bin value represents the total percent of all peptide identifications in that bin range for the corresponding method. The distributions in a) are calculated using only unique phosphopeptides identified by all three methodologies, whereas the distributions in b) are calculated using all peptides with probability >= 0.95.

To generate the figure, counts between methods were normalized to 1 to display relative bin size. The MS3-sum method produced the largest fraction of peptides in the 19+ bin amongst the MS3 methods, and was thus chosen for comparison to MS2 and MSA, and MS3-1-p. The overall results are similar in all methods, with MSA producing a higher fraction of significant (19+) localization scores. Also, the MS3-1-p shows a skew towards lower localization score values. An interesting variation arises in a manner that is dependent on the subset of peptides used in the comparison. The top panel of the figure is restricted to a comparison of peptides identified by all three methods, whereas the bottom panel utilizes the localization scores for all peptides meeting or exceeding a probability threshold of 0.95, thus allowing peptides to be included that are identified uniquely by one or two methods. The bottom panel indicates a more pronounced increase in the ratio of peptides achieving a significant localization score using the MSA method than the top panel.

It should be noted that the localization calculations did not utilize the ion assignments used by Mascot for generating a score for a peptide assignment. Our localization score calculation takes as input a peaklist and the corresponding peptide sequence assigned by Mascot. The algorithm calculates theoretical fragmentation masses based on the peptide. Fragment masses are assigned to a peak in the spectrum if a calculated fragment mass is within a user-defined threshold of that peak (the fragment mass tolerance use in the Mascot search, 0.8 Da, was used for these data). The theoretical fragmentation calculations are done using custom Python modules which are themselves based largely on a Python library written by the Pevzer lab, freely available as part of the InsPecT search platform[14] (39).

Table 4-3 shows statistics for the various ion types assigned to all high-scoring (>=0.95 probability) peptides for each method. Although an optimal peak depth is used in the calculation of the localization score as described in (31), the peak depth was fixed at four to generate the results given in the table (peak depth indicates the number of top peaks selected per 100 Da window for the purposes of scoring). Ions were assigned to

[14] http://peptide.ucsd.edu/Software/Inspect.html

peaks in a ranked order based on their likelihood, such that a peak identified by a more likely ion type would not be replaced by a less likely one. Note that even though all these ion types were annotated, not all were utilized in the calculation of the localization score. The table indicates that all methods are roughly similar in the types of ions generated, with no significant bias. Any performance improvement found by MSA in localizing a site of modification does not appear to be a function of selection of a particular set of ions. The ion percentages in the table were calculated for peptides identified by all methods, corresponding to Panel 4-4a; the identical percentages calculated using all ions for a given method (Panel 4-4b) shows very little variation (data not shown). Overall, the number of peaks identified per spectrum was very consistent across all methods as well, with the lowest method averaging 29.6 ions/spectrum and the highest 32.7 ions/spectrum.

Ion	10186-MS2	10187-MS2	10189-MSA	10190-MSA	10192-MS3	10193-MS3
M-P	1.33	1.13	0.14	0.12	1.08	1.06
a	0.72	0.93	0.86	0.95	1	0.94
a-h2o	0.81	1	0.87	0.94	0.83	1.05
a-nh3	0.98	0.65	0.95	0.57	0.85	0.95
b	18.73	18.72	17.1	17.76	15.54	15.36
b-h2o	2.8	2.71	2.45	2.3	2.3	2.34
b-h2o-h2o	0.05	0.03	0.03	0.03	0.04	0.02
b-h2o-nh3	0.05	0.07	0.03	0.03	0.05	0.04
b-nh3	2.58	2.44	2.36	2.23	2.12	2.12
b-p	10.09	9.69	11.87	12.72	11.85	12.74
b-p'	0.96	1.03	1.3	1.24	1.53	1.62
b2	6.39	7.61	6.83	6.34	5.59	5.8
b2-h2o	0.02	0.02	0	0	0	0
b2-nh3	0.08	0.2	0.18	0.09	0.06	0.11
b2-nh3-h2o	0.03	0.05	0.03	0	0.06	0.01
b2-p	1.43	1.46	1.55	1.39	1.86	1.68
b2-p'	1.13	1.23	1.12	1.26	1.37	1.04
b3	0.86	1.05	0.86	0.98	1.07	1.1
y	26.6	27.81	27.65	27.24	27.39	27.37
y-h2o	2.92	3.13	3.36	3.02	3.13	3.59
y-nh3	1.42	0.95	1.28	1.42	1.24	1.19
y-p	2.95	2.33	3.41	3.34	4.37	3.78
y-p'	0.35	0.43	0.33	0.44	0.38	0.27
y2	9.34	8.96	8.14	8.46	8.12	8.15
y2-h2o	0	0.02	0	0.01	0.01	0.01
y2-nh3	0.06	0.17	0.12	0.07	0.05	0.12
y2-nh3-h2o	0.03	0.07	0.07	0.04	0.01	0.02
y2-p	4.98	3.66	4.94	5.05	5.5	5.25
y2-p'	0.9	0.88	0.78	0.89	1	0.92
y3	1.42	1.55	1.3	1.07	1.58	1.28

Table 4-3. **Ion statistics for confident peptide identifications by methodology**. Rows indicate ion types. The numbers indicate the percentage of total assigned peaks for each method that were labeled as the given type. Ion types are sorted alphabetically. A 2 or a 3 after an ion type indicates the charge of the ion; -p indicates a 98 Da neutral loss and –p'. loss of 80 Da. M-P is a label given to the precursor neutral loss peak. Percentages for the MS3 methods were generated from the unadjusted probability results. Peptides with probability >= 0.95 were used to generate statistics.

4.4.5 Effect of method on instrument mass accuracy

As a last comparative measure, peptide assignments were queried to detect any difference in mass accuracy that may occur between the different methods. To perform this comparison, 120 high-scoring peptides identified by all three of MS2, MSA, and MS3 methodologies were examined. The root mean square (RMS) error as reported by Mascot for each peptide assignment was manually recorded (this number is available in the Peptide View page of an individual result) and plotted, shown in Figure 4-5. This number represents the overall RMS error for all theoretical assignments to peaks in the experimental spectrum. As noted earlier, database searches were performed both with and without internal ions; the inclusion of internal ions resulted in a significant increase in the average RMS error rate, not unexpected given the high rate of false peak assignments for these ions. All results reported in this figure do not include internal ion masses in the mass error calculation.

The RMS error results are matched vertically by instrument method for each of the individual unique peptide sequences and sorted by peptide length along the x-axis. For instance, for Peptide ID #1, the blue circle (MS2), red square (MSA) and green triangle (MS3) for the seven-mer sequenced peptide ADSLIYK is shown clustered together around 200 ppm RMS. This peptide had a calculated RMS mass error of 188, 192 and 199 ppm for each of the methods, respectively. The last data points along the x-axis represent a peptide of length thirty-nine, which produced a calculated RMS mass error of 505, 549 and 410 ppm for MS2, MSA, and MS3 instrument methods, respectively.

While these two examples show MS2 as having better mass accuracy than MSA, linear regression curves indicate that MSA spectra have an overall better mass accuracy than the other two methods. As can be seen from the figure, MSA identifications are roughly 30 ppm better than MS2 and 50 ppm better than MS3 spectra over much of the range of these different peptide lengths. As the overall length of the peptide increases, the RMS error for the three methods approaches a similar value. A normalization effect occurs as the length, and thus the number of identified ions, increases. This overall trend

of MSA having better mass accuracy than the other methods is interesting: one might expect MSA to have a higher mass error due to space charge effects within the ion trap. During MSA analysis, relative to MS2 and MS3, the ion trap is left open longer without interludes of evacuation. Therefore, the greater number of ions collected in the MSA method compared with MS2 and MS3 would increase space charge effects and thus decrease mass accuracy.

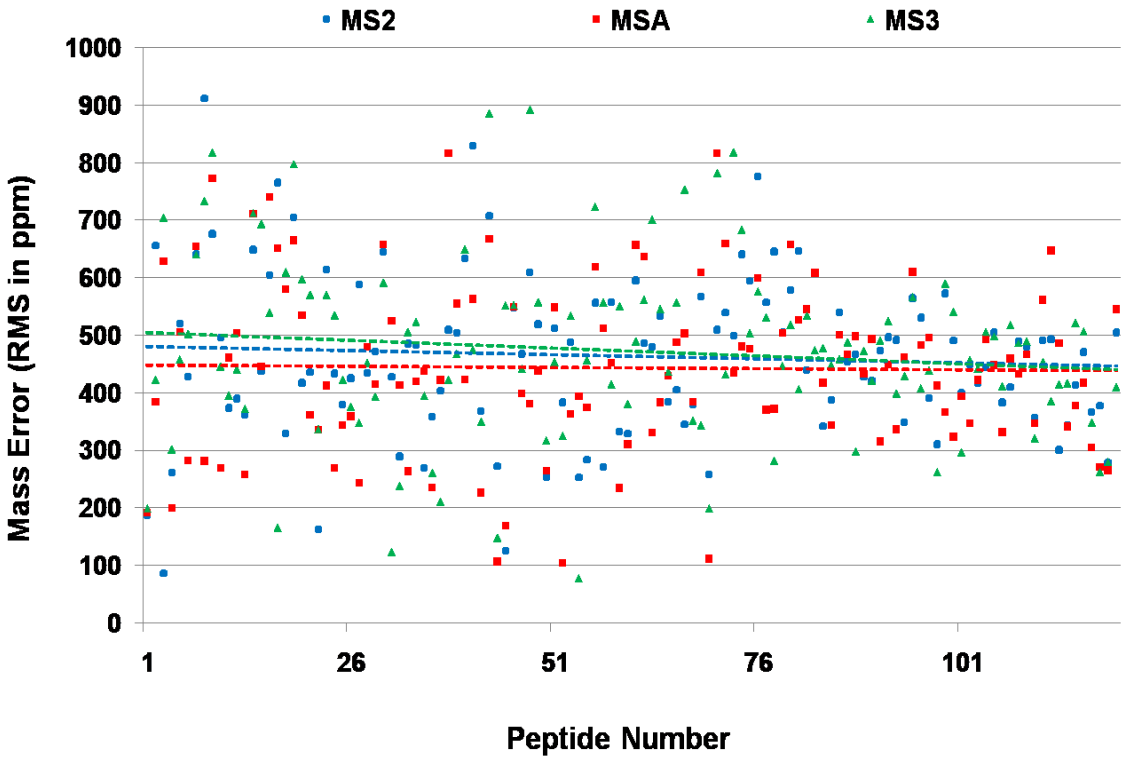


Figure 4-5. Mass accuracy of fragment ion assignments from MS2, MSA, and MS3 methodologies. Results for individual peptides are sorted on the horizontal axis by decreasing mass error. The vertical axis shows the RMS mass error in ppm of theoretical fragment ion matches to the experimental spectra produced by MS2 (blue circles), MSA (red squares), and MS3 spectra (green triangles). Linear regression curves are fit to the data for each method, shown as dashed lines blue (MS2), red (MSA) and green (MS3).

4.5 Conclusions

In the methodologies described, a complex phosphopeptide-enriched sample was analyzed in replicate using several different run methods on an FT-LTQ mass spectrometer. At one level, the methods generated a similar number of identifications: the difference in the number of unique peptide assignments between the best- and worst-performing algorithms is 8%. The resulting MS2 datasets contained the largest number of MS^2 spectra, although the MS3 datasets contained more overall MS^2 and MS^3 spectra combined. However, given that the net intensity of MS^3 spectra is significantly lower than MS^2, one might expect the net result to be a higher number of peptide identifications from MS^2 spectra even for phosphopeptide-enriched samples. The fact that current instrumentation often generates sufficient fragmentation information to uniquely identify a phosphopeptide despite the dominance of the neutral loss peak supports this notion. On these data, however, this was not the case. Out of the three methodologies discussed, MS2 produced the fewest number of unique peptide identifications. The most overall (phosphorylated + unmodified) modifications were found by the MSA methodology and the MS3 methodology that utilized post-search probability correction and the 1-(1-P_{MS2})(1-P_{MS3}) score: the MS3-1-p method. More total phosphorylated peptides were identified by the MS3 methods, specifically the probability-corrected (MS3-ADJ and MS3-1-p) approaches. Although these two approaches produced the most overall phosphopeptide identifications, the MS3-sum method resulted in a higher number of localized phosphorylation sites than the other MS3 methods. The overall highest number of significantly localized sites of phosphorylation, however, was found by the MSA method. MSA produced the fewest number of spectra but generated the largest fraction of successfully identified spectra. MSA spectra appeared to have a slight advantage in localizing the site of modification as demonstrated by a minor positive skew in the overall Ascore distribution, and also exhibited a detectably lower overall mass error.

The overall result of the above analysis confirms that MSA is a convenient and effective overall approach. Given the fact that MSA spectra are simple to utilize in that they do not require non-standard database searching or additional MS^2/MS^3 combination approaches and bookkeeping, they appear to be an optimal methodology, at least for the

automated processing pipeline I describe here. A different biological sample, perhaps one generated by a different phosphopeptide enrichment strategy that produces different ratios of multiply-phosphorylated forms, or a less complex sample, may produce different results. Other instrument conditions will likely have an effect on the utility of MS^3 data as well. The fact that both the MSA and the MS3 methodologies performed better than the MS2 methodology suggest that there is additional useful information contained in MS^3-type data that can be utilized by a standard automated search engine in a high-throughput approach. A manual curation strategy, or more directed algorithms focused specifically on identifications of phosphorylated peptides, could produce a different set of outcomes on these data.

References

1. Hunter, T. (2000) Signaling--2000 and beyond. *Cell* 100:113-27.

2. Cohen, P. (2000) The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends Biochem. Sci.* 25: 596–601.

3. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* 298, 1912-34.

4. Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., Manning, G. (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A.* 101, 11707-12.

5. Kersten, B., Agrawal, G. K., Iwahashi, H., Rakwal, R. (2006) Plant phosphoproteomics: a long road ahead. *Proteomics* 6: 5517-28.

6. Mackay, H. J., Twelves, C. J. (2007) Targeting the protein kinase C family: are we there yet? *Nat Rev Cancer* 7: 554-62.

7. Rikova, K., et al. (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* 131: 1190-203.

8. Guo, A., et al. (2008) Signaling networks assembled by oncogenic EGFR and c-Met. *Proc Natl Acad Sci U S A.* 105: 692-7.

9. Mazanetz, M. P., Fischer PM. (2007) Untangling tau hyperphosphorylation in drug design for neurodegenerative diseases. *Nat Rev Drug Discov*. 6: 464-79.

10. Hoffert, J. D., Knepper, M.A. (2008) Taking aim at shotgun phosphoproteomics. *Anal Biochem.* Nov 22: [Epub ahead of print]

11. Witze, E. S., Old, W. M., Resing, K. A., Ahn, N. G. (2007) Mapping protein post-translational modifications with mass spectrometry. Nat Methods. 4: 798-806.

12. Ptacek, J., Snyder, M. (2006) Charging it up: global analysis of protein phosphorylation. *Trends Genet.* 22: 545-54.

13. Pflieger, D., Jünger, M., Müller, M., Rinner, O., Lee, H., Gehrig, P., Gstaiger, M., Aebersold R. (2007) Quantitative proteomic analysis of protein complexes: Concurrent identification of interactors and their state of phosphorylation. *Mol Cell Proteomics*. Oct 23 [Epub ahead of print]

14. Bodenmiller, B., Mueller, L. N., Pedrioli, P. G. A., Pflieger, D., Jünger, M. A., Eng, J., Aebersold, R., and Tao, W. A. (2007) An integrated chemical, mass spectrometric and computational strategy for (quantitative) phosphoproteomics: Application to Drosophila melanogaster Kc167 Cells. *Molecular BioSystems*, 3, 275-286.

15. Villén, J., Beausoleil, S. A., Gerber, S. A., Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A.* 104: 1488-93.

16. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 4, 310-327.

17. Ballif, B. A. Villén, J., Beausoleil, S. A., Schwartz, D., Gygi, S. P. (2004) Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* 3: 1093–1101.

18. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., White, F. M. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisae. *Nat. Biotech.* 20, 301–305

19. Tholey, A., Reed, J., Lehmann, W. D. (1999) Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J Mass Spectrom*. 34: 117-23.

20. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 101, 13417-13422

21. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 101, 12130-12135

22. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell.* 127: 635-48.

23. Macek, B., Mijakovic, I., Olsen, J.V., Gnad, F., Kumar, C., Jensen, P. R., Mann, M. (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium *Bacillus subtilis. Mol Cell Proteomics* 6: 697-707.

24. Hoffert, J. D., Wang, G., Pisitkun, T., Shen, R.F., Knepper, M. A. (2007) An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins. *J Proteome Res.* 6: 3501-8.

25. Ulintz, P. J., Bodenmiller, B., Andrews, P. C., Aebersold, R., Nesvizhskii, A. I. (2008) Investigating MS^2-MS^3 matching statistics: A model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. *Mol Cell Proteomics* 7, 71-87.

26. Wu, J., Shakey, Q., Liu, W., Schuller, A., Follettie, M. T. (2007) Global profiling of phosphopeptides by titania affinity enrichment. *J Proteome Res.* 6: 4684-9.

27. Palumbo, A. M., Tepe, J. J., Reid, G. E. (2008) Mechanistic Insights into the Multistage Gas-Phase Fragmentation Behavior of Phosphoserine- and Phosphothreonine-Containing Peptides. *J Proteome Res.* Jan 9; [Epub ahead of print]

28. Li, X., Gerber, S. A., Rudner, A. D., Beausoleil, S. A., Haas, W., Villén, J., Elias, J. E., Gygi, S.P. (2007) Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J Proteome Res.* 6, 1190-7.

29. Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F., Coon, J. J. (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem*. 76: 3590-8.

30. Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 20: 3551-67.

31. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*. 24: 1285-92.

32. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., and Aebersold, R. (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nature methods* 4, 231-237.

33. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1, 2005.0017.

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

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

36. Bairoch, A., et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33:D154-159.

37. Elias, J. E., Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. Mar 4: 207-14.

38. Käll, L., Storey, J. D., Maccoss, M. J., Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res.* 7: 29-34.

39. Tanner, S., Shu, H., Frank, A., Wang, L., Zandi, E., Mumby, M., Pevzner, P. A., Bafna, V. (2005) InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal Chem.* 77: 4626-39.

Chapter 5

Differential expression of a dengue infected human cell line using four different fractionation methods

5.1 Introduction

One of the results often observed in large-scale proteomics studies is the degree to which different analyses of a similar or identical biological sample produce different peptide identifications. The funding of large, multi-group projects such as the HUPO Plasma Proteome Project (PPP) (1), and the recent availability of public repositories of protein mass spectrometry data, provides an opportunity for large inter-group data comparisons that explore this observation. In a current publication, Klie et al. present results in which they use natural language processing techniques to look for latent semantic patterns in large datasets (2). The analysis produces measures of similarity between experiments based on the number of proteins identified between them, and permits a view of overall similarity patterns across all the data. The results demonstrate a dramatic lack of reproducibility between experiments. Not a single protein is identified in all of the submitted datasets of the HUPO plasma sample, 95 experiments total. Out of the 7884 proteins identified, only 40 are identified in more than half the experiments, and 70% of the reported protein identifications are found in only one or two experiments.

The results can be partially attributed to the variety of experimental approaches and instrumentation utilized, and also to the well-known under-sampling issue common to LC-MS/MS experiments. The study confirms the latter finding, indicating a much higher degree of similarity between 2D gel-based experimental methodologies than LC-based shotgun methodologies.

Other researchers have systematically looked at the degree to which replicate analyses yield new unique identifications. In examining the data submitted to the PeptideAtlas database (3), the Aebersold group observes that the rate at which new unique peptides are identified can begin to saturate due to overlapping identifications from similar experiments. The saturation occurs well below the threshold of the total number of predicted peptides for a genome, and appears to be more of a function of experimental methodology and instrument. In yeast, although still below a complete saturation, they identified a total of 36133 unique peptide identifications with PeptideProphet scores > 0.9 from 4.9 million MS^2 spectra (the number included in PeptideAtlas at the time of publication of the manuscript) (4). The calculated number of detectable (mass range 500 to 4000 Da) tryptic peptides in yeast, by contrast, is 436,445. All methodologies combined were thus only identifying less than 10% of the theoretical tryptic peptides in the genome.

These results have obvious implications for experimental design. The study in (2) suggests that if the goal of an experiment is to produce a broad coverage of all proteins in a sample, a shotgun methodology might be more appropriate; whereas for a quantitative study of a simpler protein mixture reproducibility might be a priority, suggesting a gel-based methodology. The PeptideAtlas evaluation suggests that introduction of different experimental methodologies can introduce new peptide identifications at a faster rate. Also, different computational processing strategies applied to the same dataset can produce different and complementary results as well. Using multiple search engines, for example, has been demonstrated to be quite useful in selecting more peptide identifications (5).

Over the course of several years, an analysis of differential expression of proteins as a result of viral infection has been generated by our group. The overall goal of the study was to examine the differential expression of proteins in a human cell line as a result of infection with the dengue virus. The sample is a challenging one in that the virus is somewhat subversive: it hijacks the transcriptional machinery of the cell in a subtle manner and does not outright kill the cell in a short period of time. Infected cells are harvested 48 hours after infection, at which point there is enough time for a significant viral titer to accumulate in the cell, but not a sufficient amount to destroy it. To be able to detect significant expression changes, a more involved proteomics strategy that permits a deeper view into the proteome of the human cell was thus necessary. Overall, differences in the relative expression of several significant proteins were found in the course of the study. These biological findings will be presented elsewhere. The goal of this portion of the thesis is to use the datasets generated in this study for computational analysis.

The data generated in the course of this project are illustrative for examining issues of depth and breadth of proteomics experiments, and the degree to which various experimental approaches are 'orthogonal'—yield unique identifications— and complementary. There are several bases for method differentiation that may be examined in these data: intact vs peptide-level iTRAQ labeling; four methods of fractionation: strong cation exchange (SCX), peptide isoelectric focusing (IEF), polyacrylamide gel (SDS-PAGE), and size-exclusion chromatography; and different search engine analyses. What follows is a comparison of these methodologies, focusing on the degree to which different methods produce orthogonal results and, in essence, justify variation in method when the goal is to identify a larger number of lower-abundant proteins.

5.2 Experimental Methodology

Sample Preparation: The samples utilized for the study consisted of $1 \cdot 10^8$ U937 human monocytic cells. Two full samples were utilized, one a control and one infected with Dengue 2 New Guinea C strain for 48 hours. Samples were washed and lysed enzymatically with the Sigma CelLytic-M Lysis solution with added mammalian protease and phosphatase inhibitor cocktails and 200 mM Tris-2-carboxyethyl phosphine (TCEP) in 1.5M Tris-HCl, pH 8.8.

iTRAQ labeling: Samples were labeled with the iTRAQ reagent in two ways: at the peptide level following digestion and intact, at the protein level. At the peptide level, after an acetone precipitation, the protein precipitate was resuspended, reduced, and alkylated with the reagents contained in the iTRAQ kit exactly as described in the manufactures instructions (Applied Biosystems, Foster City, CA). Proteins were then digested overnight with trypsin and labled following the standard iTRAQ protocol. The intact labeling protocol is almost identical, except there was no trypsin digestion: the iTRAQ labeling reagent was added directly to the protein solution and the mixture was incubated at room temperature for 2 hours.

1st dimension fractionation: Peptide iTRAQ labeled samples were separated via two means of fractionation: strong cation exchange (SCX) and isoelectric focusing (IEF). For SCX, peptides were loaded onto polysulfoethyl-A spin columns (SEM HIL-SCX, PolyLC, The Nest Group. Southboro, MA). The bound peptides were washed then eluted in a stepwise gradient of increasing salt concentration (50, 80, 115, 155, 180, 205, 350, and 500 mM of KCl) in equilibration buffer. For IEF, the OFFGEL in-solution IEF fractionation device was utilized (Agilent Technologies, Santa Clara, CA). 24 cm pH4-7 IPG strips were used in the device to fractionate samples into 24 fractions. Resulting fractions were cleaned on a C_{18} column (Sep-Pak Vac 1cc, Waters). Collected fractions from both methods—eight for SCX and 24 for IEF—were dried in a vacuum centrifuge and reconstituted with 43 µl of 0.1% TFA in water for reverse phase chromatography.

Protein-level iTRAQ samples were separated using two different means as well: SDS-PAGE and size-exclusion chromatography (SEC). For SDS-PAGE separation, the

dry, labeled protein samples were reconstituted in gel loading buffer and separated with a 1 mm 4-20 % Nu-PAGE Bis-Tris gradient gel (Invitrogen, Carlsbad, CA). The gel was cut into 11 sections and the excised gel bands were digested overnight with with trypsin. For SEC, the iTRAQ labeled protein sample was placed on a Zorbax GF 250 4 μm (4.6 x 250) column (Agilent Technologies, Santa Clara, CA). One minute fractions were collected across the elution profile (750 μL/min flow rate) for a total of eight fractions. These fractions were concentrated in a speed vacuum centrifuge, resuspended, and subjected to trypsin digestion. Following digestion, eluted peptides were dried in a vacuum centrifuge and reconstituted in 0.1% TFA for reverse phase chromatography.

Mass spectrometry: 1st dimension fractions from all four methods were reverse phase (RP) separated by C_{18} nano LC and spotted onto a stainless steel Opti-TOF™ MALDI Plate System with a spotting robot. Samples were spotted at 10-second intervals, 384 spots per plate. The spotted fractions were analyzed in the ABI 4800 TOF/TOF mass spectrometer. First stage MS analysis was completed in positive ion, reflector mode acquiring precursor ions in a mass range of 950-4000 *m/z* for protein level labeling and 800 – 3500 *m/z* for peptide level labeling. Tandem MS analysis was completed in a data dependent manner in which the most abundant 15 and 8 peaks were selected per spot, with a minimum S/N 40 and 100, for protein level labeling and peptide level labeling, respectively. Fragmentation of all peptides was induced by the use of atmosphere as a collision gas with collision energy of 1kV.

Database Searching: Searches were conducted using four algorithms: SEQUEST (version 27, rev. 12) (6), Mascot (version 2.1.0) (7), X!Tandem (version 2006.09.15.1) (8), and Protein Pilot with the Paragon algorithm (version 2.0) (9). The database utilized by all search algorithms was generated by concatenating the reversed protein sequences to the forward sequences of the entire human IPI database (version 3.29, updated May 2007). Where indicated and specified, database searches consisted of using trypsin enzyme specificity, a mass tolerance of 0.7 Da on parent ions and 0.3 Da on fragment ions, a maximum of two missed cleavages. Variable modifications included methylmethane thiosulfonate (MMTS) and iodoacetamide on cysteine, iTRAQ reagent on lysine, tyrosine, and the peptide N-terminus, deamidation of asparagine and glutamine,

and oxidation of methionine. Paragon does not utilize these database search parameters, but uses an unrestrained search algorithm that is detailed in (9). Proteins in ProteinPilot were accepted as confident identifications if the probability ≥ 95% probability as calculated by Paragon. Scaffold (version 01_06_03)[15] was used to visualize MS/MS based peptide and protein identifications from SEQUEST, Mascot and X!Tandem. Identifications were accepted if they exceeded the 95% protein probability and a 90% peptide probability calculated by the Scaffold implementation of the Prophet algorithms (5, 10, 11) Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

5.3 Search Engine Comparison

Processing the results for all four methodologies provided an opportunity to assess the performance of the various mass spectrometry database search engines, and to evaluate a new, promising search engine: ProteinPilot/Paragon (6). Paragon relaxes normally specified constraints on the mass tolerance of identifications and automatically searches a large number of possible PTMs. It has been known for some time that search engines produce different results on identical datasets: after all, their scoring algorithms are not identical and function to reward attributes in the data in different ways. Typically the differences in identification tend to be "borderline" cases, with the most obvious examples of both correct and incorrect assignments assigned correctly by all tools. The size of this "grey area" differs between search engine comparisons, however, so it is instructive when using a new algorithm to evaluate the relative overlap between its results and the results of other algorithms.

An excellent tool for integrating the results of the "Big Three" most commonly-used search engines—SEQUEST, Mascot, and X!Tandem—has been developed: the Scaffold tool by Proteome Software. This tool takes as input results from these three engines and calculates new probability scores for each peptide and protein assignment

[15] http://www.proteomesoftware.com/

using a customized implementation of the Peptide and ProteinProphet tools. A new "agreement score" method is also implemented based on evidence from all algorithms for a match. Overall, the software boosts the probabilities for peptides identified by multiple engines, and effectively permits an increase in sensitivity as a result (5).

Scaffold was used to integrate the results from the Big Three search engines as a basis for comparison with ProteinPilot. The results from all four searches are indicated in Figure 5-1. The results indicate several interesting observations. Overall, Paragon identifies many more proteins than any of the other methods. Many of the proteins that Paragon identifies are also identified by one or more of the other search tools, but there are 444 unique proteins found by Paragon alone. The identifications from the other three search engines produce relatively similar results in that the number of identifications common to all three is the largest area.

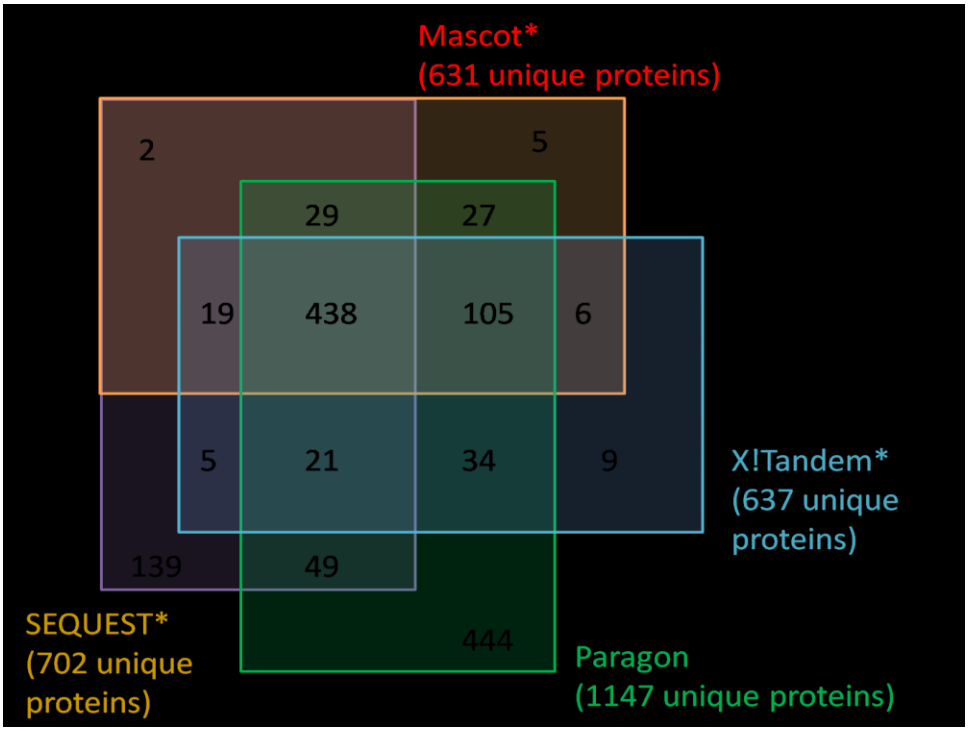


Figure 5-1. Comparison of the number of protein assignments between search engines. Numbers represent total unique protein identifications meeting the 0.95 probability threshold as defined by the search engine; searches marked with '*' were thesholded via Scaffold data integration software. Results represent data from all four fractionation methodologies. The Venn diagram is not area-proportional.

These results suggest an orthogonality in the way Paragon scores that the other engines do not share. We know this to be true theoretically in that Paragon implements a somewhat "blind", unrestricted search for a large number of PTMs, and also permits the inclusion of peptides with significantly more broad delta mass errors under defined circumstances (9). The results, however, require a justification if the results from Paragon are to be believed: it is a new algorithm producing results that have not had the benefit of years of validation on the behalf of pundits in the community. Since Scaffold provides a mechanism to integrate the results of the three other, more established algorithms, it may be used as a basis of comparison with Paragon. Toward that end, the integrated results from Scaffold were compared with the Paragon algorithm at both the protein and peptide levels (Figure 5-2).

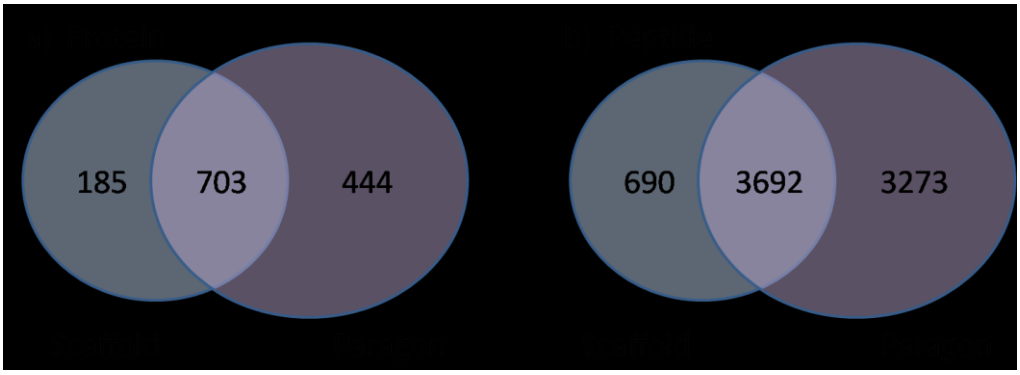


Figure 5-2. Comparison of Scaffold and Paragon results at the protein and peptide levels. Areas show unique identifications; protein identifications are based on unique gene name, peptide identifications based on unique sequence. To be included in the results, protein probabilities must meet the 95% confidence threshold and include two or more unique peptide identifications.

Figure 5.2a presents similar data as the previous figure, but with the results from the Big Three engines integrated. As can be seen from Figure 5-2b, 43% of the total peptide identifications are identified by Paragon alone, compared to 9% with the other engines. If such results can be confirmed, these data indicate a significant enhancement in performance by Paragon above the other search algorithms. Nevertheless, there is

merit in running the others in that a significant number of identifications were made by them that were not found in Paragon.

Based on an examination of the unique Paragon identifications, the majority of these results are explainable as being based on features not accommodated by the other search tools. The numbers of unique Paragon IDs having specific features are listed in Table 5-1. Note that the percentages are not cumulative; an individual peptide identification can have more than one of these features. 62% of the unique Paragon identifications corresponded to non-standard tryptic cleavage. The extensive number of additional modifications calculated by Paragon that are not accommodated by standard search approaches accounted for at least 12%. A number of other less common modifications were found in a smaller number of peptides, not shown in the table. A large portion of identifications, 5%, could be attributed to iTRAQ labeling of non-standard amino acids Ser and His. Surprisingly, only 2% of the unique identifications were found to have precursor mass errors > 0.7 Da. Given the range of the timed ion selector (TIS) on the 4800, a larger number was expected. A significant number of identifications with > 2 missed cleavages were identified as well.

Several of these results have been confirmed using other search engines. Promiscuous iTRAQ labeling, for example, is a known event. Confirmation searches in which the Mascot search engine was specifically programmed to find abnormal labeling of Ser and His produced positive results. Loosening the search constraints of the standard tools to accommodate partially tryptic search would pick up a number of these identifications in the standard tools, at the expense, however of an order of magnitude larger search time and an increase in the false positive rate. A number of abnormal modifications have been examined by expert curators as well and have been found to be believable. Taken together, these results are an indication that the additional Paragon peptide assignments are largely valid. Based on this premise, the bulk of the work in the remainder of this study utilize Paragon results only for simplicity of articulation.

Non-standard feature	Percent of unique Paragon identifications
Atypical (non-tryptic) cleavage: N-term	45
Atypical (non-tryptic) cleavage: C-term	17
Atypical iTRAQ labeling (S, H residues)	5
More than 2 missed cleavages	4
Acetaldehyde modification	3
Atypical Oxidation (P, H residues)	3
Delta mass > 0.7 Da	2
Formyl/Carbamyl@N-term	1

Table 5-1: Frequency of non-standard features amongst unique Paragon identifications. Feature percentages are not exclusive; a peptide might be counted in the total of more than one feature.

5.4 Comparison of Fractionation Methodologies

The identifications from the various fractionation methods are now compared. The results presented here represent a comparison of four different methods of 1st dimension separation, two using digests of intact iTRAQ-labeled peptides and two using standard peptide-level iTRAQ-labeled peptides: All experiments consisted of two experimental replicates each of uninfected control and viral infected sample. The uninfected samples were labeled with the 114 and 115 iTRAQ labels and the infected the 116 and 117 labels, and the four samples mixed. The two protein-level fractions are:

- Protein-level size-exclusion (Protein-SE) fractionation: 100 μg each sample, 400 μg total loading.
- Protein-level SDS-PAGE (Protein-Gel): 25 μg each sample, 100 μg total loading.
- Peptide-level isoelectric focusing (Peptide-IEF): 100 μg each sample, 400 μg total loading.

- Peptide-level strong cation exchange (Peptide-SCX): 50 µg each sample, 200 µg total loading.

Fractions from each 1st dimension separation were further fractionated by reverse phase chromatography for analysis as described in the methods section. All results were processed using the Protein Pilot/Paragon software and compared. The number of unique identifications obtained by each of the methods, and their relationships, are shown in Figure 5-3.

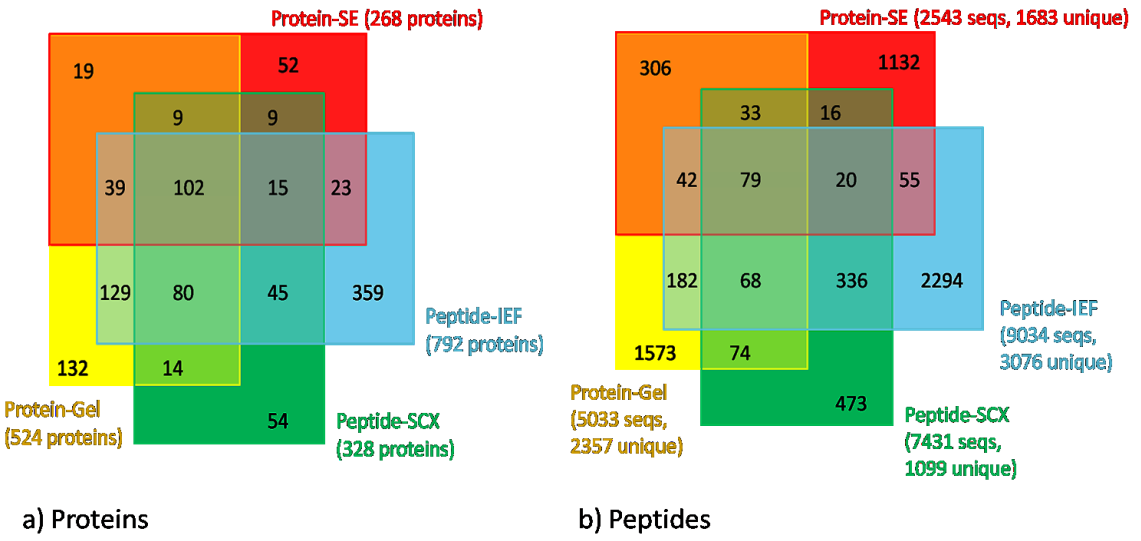


Figure 5-3. Comparison of unique peptide and protein identifications between four fractionation methodologies as identified by the Paragon algorithm. Results including if P > = 0.95; protein identifications required a minimum of 2 unique peptides.

Results at the peptide level (Figure 5-3b) indicate a surprisingly small overlap between the identifications by the various fractionation methods. In all, 80% of the peptide identifications were seen by only one method. In contrast, only 79 peptides,1.2%, were identified by all four methodologies. A small overlap size could be expected between the intact vs peptide-level iTRAQ, in that the protein level labeling would prevent tryptic digestion at lysines. However, a quite significant disparity exists between the methods at the same levels of iTRAQ labeling; between the protein-labeled levels, for example, the vast majority of the identifications were of different peptides.

The methods do vary quite largely in the number of identifications returned. The largest number of identifications results from the Peptide-IEF method, followed by the Protein-Gel method. The variance can be attributed to both sample loading and number of 1st dimension fractions to some degree, although not completely so. Peptide-IEF and Protein-SE both had the highest loadings at 400 µg, whereas the Peptide-SCX and Protein-Gel had 200 µg and 100 µg total loading, respectively. Peptide-SCX and Protein-SE both had only eight fractions, whereas Protein-Gel had 11 and Peptide-IEF had 24. It is therefore not terribly surprising that the most successful experiment, Peptide-IEF, had both the highest loading and the largest number of fractions. Its notable that the Protein-SE experiment had by far the lowest overall total number of significant identifications (designated 'seqs' in the figure), but had more unique identifications than the Peptide-SCX method. This can likely also be attributed to the intact labeling, yielding overall longer peptides with more information content.

At the protein identification level (Fig 5-3a), the same pattern is observed, however there is a degree of convergence in the identifications: 9% of the proteins are identified by all four methods and only 55% are identified uniquely by one method. This still represents a significant lack of overlap, however. Protein-SE generates the fewest number of unique protein identifications, with Peptide-IEF generating the most at the protein level. Overall, the intuitive expectation might be that the same proteins should be identified by all methods, indicating that any one method is effectively sampling the protein complement of the sample. These results provide a different picture, however, and suggest that different fractionation approaches are indeed complementary and orthogonal, and each contributes uniquely to the final result.

5.5 Cumulative Fraction Orthogonality

When processing total results, each fraction from any methodology was spotted separately onto MALDI target plates and analyzed in the mass spectrometer. These results were extracted individually, and provide another instructive dimension of analysis. In the case of the Peptide-SCX experiment, fractions were spotted multiple times to examine the effect of sampling rate and saturation on identical samples. Also, in that the

SCX methodology was done using spin columns, flow-through fractions were spotted and analyzed to identify peptides that failed to bind to the column. The net result of this methodology was a set of individual data sets that could be examined individually for uniqueness.

Figure 5-4 presents the cumulative number of unique peptide identifications as a function of number of "quality" spectra acquired, these being defined as spectra producing a hit to a sequence with a probability P >= 0.95. Peptide identifications were counted in a cumulative manner, adding to the total if a unique new peptide is identified. The largest number of fractions that were analyzed in the mass spectrometer was from the Peptide-SCX experiment. The number of unique identifications produced by this method appears to saturate, likely due to replicate spottings of identical fractions. This method also produced the least number of significant identifications, overall suggesting that increasing the number of salt elutions in this experiment might not be the most productive endeavor.

The other methods show a more distinctly linear gain; the number of unique identifications by fraction appears to be consistent, with minor variation by fraction. An exception is the Protein-SEC experiment, which shows several fractions stacking near each other on the plot. This result is a function of the specifics of the way the samples were spotted on the plate; the number of spots was much fewer in several of the fractions. In all, the non Peptide-SCX fractions seem far from saturation, and increasing fraction number or repetition of these experiments would likely result in a number of new identifications.

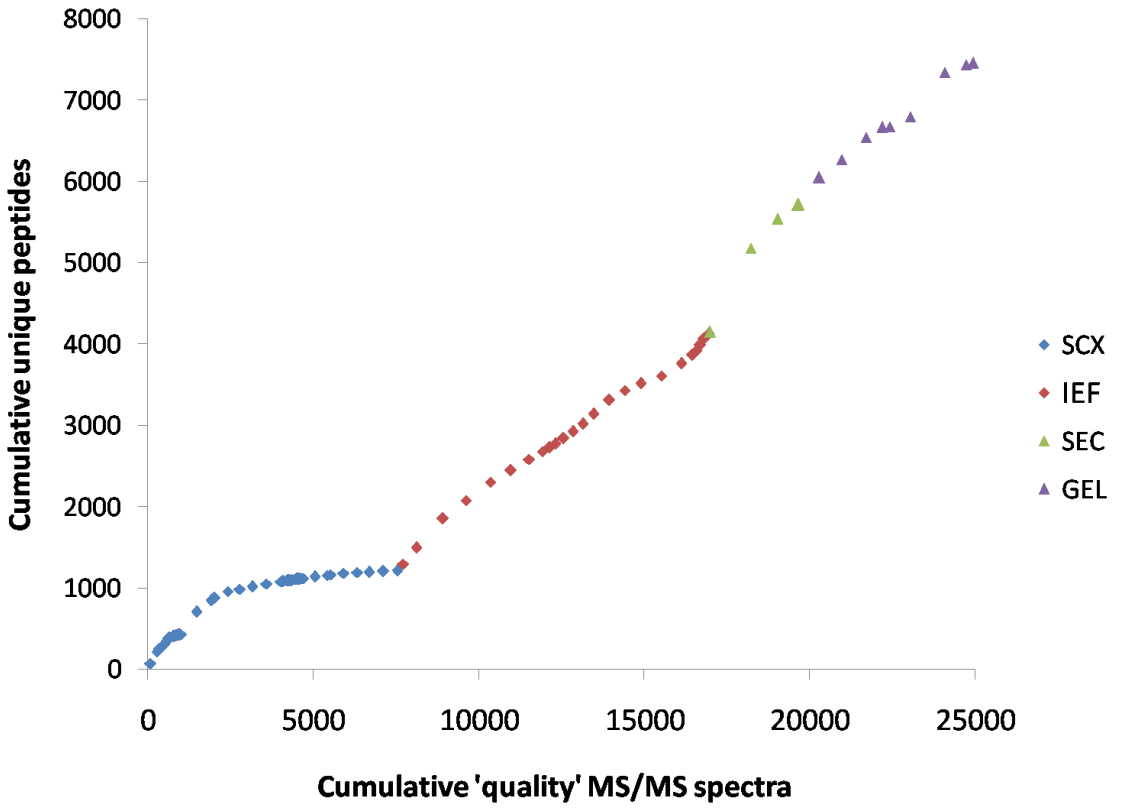


Figure 5-4. Cumulative number of MS/MS spectra versus number of unique peptide identifications by fraction. Fractions are ordered by method, with peptide-level iTRAQ methods to the left and protein-level on the right. Peptides included with P >= 0.95, and spectra included if they produce a significant peptide.

5.6 Conclusions

The results of this analysis show a surprisingly small amount of overlap between unique peptides identified by the four different experimental methodologies. The methods are thus quite complementary to one another in terms of providing a significant number of identifications per number of spectra. The one experiment that was replicated twice, the SCX method, did not seem to see a significant benefit from the repetition. It was, in fact, the only method that showed saturation in the cumulative curve. This may be partially due to the experimental fidelity of the method: SCX spin columns may simply not provide the best overall fractionation. It might also suggest, however, that

rather than choosing to replicate a particular experiment using the same methodology, a different methodology might provide a better breadth of identifications. The other fractionation methods generated a somewhat linear increase in the number of unique identifications per fraction, perhaps suggesting that the physical properties of peptides being exploited by the fractionation technique are more discriminating, providing a better resolving power. Overall, such conclusions will require much larger datasets. The establishment of public repositories of mass spectrometry should facilitate this type of analysis.

The database search engines were also demonstrated to produce complementary results. This is a well-known property of these tools, and forms the basis of the Scaffold program. The Paragon algorithm was demonstrated to produce a large amount of unique peptides not identified by the standard Big Three search engines, largely due to its novel approach to searching and its consequent ability to loosen constraints on enzyme specificity, post-translational modification, and mass accuracy that aren't currently possible in the standard tools.

This research demonstrates that MS/MS spectra from several different experimental methodologies can be jointly processed and combined to create a large dataset containing more identifications than likely would have been available from any one experimental method alone. As many more large-scale datasets become publically available in common formats that are amenable to integration, overall properties of the combined dataset can be examined which will hopefully be informative in experimental design.

References

1. Omenn, G. S., et al. (2005) 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 Aug;5(13):3226-45.

2. Klie, S., Martens, L., Vizcaíno, J. A., Côté, R., Jones, P., Apweiler, R., Hinneburg, A., Hermjakob, H. (2008) Analyzing large-scale proteomics projects with latent semantic indexing. *J Proteome Res*. 7: 182-91.

3. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res.* 34(Database issue): D655-8.

4. King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., Eng, J., Desiere, F., Flory, M., Martin, D. B., Kim, B., Lee, H., Raught, B., Aebersold, R. (2006) Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biol*. 7: R106.

5. Searle, B. C., Turner, M., Nesvizhskii, A. I. (2008) Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies. *J Proteome Res.* 7: 245-53.

6. Eng, J.K., McCormack, A.L.,Yates, J.R., III. (1994) An approach to correlate tandem mass spectral data pf peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5: 976-989.

7. Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 20: 3551-67.

8. Craig R, Beavis RC. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 20: 1466-7.

9. Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A., Tang, W. H., Keating, S. P., Hunter, C. L., Nuwaysir, L. M., Schaeffer, D. A. (2007) The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics* 6:1638-55.

10. Keller, A., Nesvizhskii, A. I., Kolker, E. Aebersold, R.. (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem*. 74: 5383-5392.

11. Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem*. **75**, 4646-58.

Chapter 6
Discussion

6.1 Summary

As a discipline, the field of Proteomics is showing concrete signs of maturing. The subfield of Proteome Informatics has now arguably reached the critical mass of research which permits it to be distinguished as a discipline of its own right. Aspects of Proteome Informatics itself are maturing as well: the most fundamental contributions of the field, the mass spectrometry database search platforms such as SEQUEST, have been in development for fifteen years. Yet despite ongoing effort, these platforms have not yet had the success necessary to allow them to be used in a completely automated manner. Many, if not most, of the spectra generated to this day go un-interpreted by a standard search engine. Many more require manual interpretation. The assignment of a post-translational modification to a peptide, for example, is still tenuous enough to require a fully-annotated spectrum to be supplied to be accepted in the major journals (1).

The focus of this research has been to extend current approaches for proteome analysis, developing a greater penetration into the proteome by increasing the number of identifications attainable in particular instances. New machine learning algorithms were explored which provided insight into the types of information that is most useful for discrimination of correct peptide identifications. Also, in the context of phosphoproteomics, new methods were introduced for transforming additional information available by coupling consecutive MS^2 and MS^3 data into a probability score.

Data mining strategies, and the implementation of a localization score for phosphopeptides, allowed an in-depth examination of mass spectrometry methodologies themselves for phosphoproteomics, addressing the question of which strategy is the most optimal. Also, an assessment of the orthogonality of various fractionation methodologies was described, demonstrating that greater proteome coverage could be achieved by utilizing multiple fractionation approaches and search engines.

In general, the methods discussed are complementary to, and indeed extend upon, now standard methods in the field which were developed to efficiently utilize additional information that may be available in a proteomics study. These algorithms, exemplified by the Prophet tools and the Scaffold software[16], can be thought of as extendable frameworks for accommodating this often easily obtainable and under-utilized additional data. In the discussion that follows, additional examples of using these types of approaches are discussed. The main results from the primary body of work in this thesis are also reviewed, with an eye towards their potential biological impact

6.2 A comment on isoelectric point and the use of pI for phosphopeptide enrichment

An excellent example of the type of information that might be additionally available in a proteomics study (and which may function as a potential discriminator) is the isoelectric point (pI) of a peptide or protein. pI is defined as the pH at which the amino acid sequence has a net charge of zero, and it can be calculated from the primary sequence of a peptide. The ability to separate peptides based on pI and to also accurately predict a peptide pI thus provides a powerful tool for shotgun proteomics; this information can provide an extra filter for false positive rate calculations in that truly random false positive matches would be expected to produce correspondingly random pI measurements. One might use this information to select peptide results that fall below normal score thresholds in the results of mass spectrometry database search algorithms,

[16] http://www.proteomesoftware.com/index.html

resulting in more peptide identifications, or increased confidence in "borderline" matches.

PeptideProphet can accommodate pI information into a probability score for peptide identification. When the pI feature is turned on in this software, the tool assumes that the input sample consists of a focused set of peptides. The software calculates the average pI of all peptides, and refines the probability score for each peptide based on how close or far the calcuated pI of the peptide is from the average (2).

A significant amount of work has been performed in our group surrounding the isoelectric focusing of peptides. One major endeavor has been the goal of utilizing pI to enrich for phosphopeptides. The following theoretical experiment illustrates the idea. The pIs of a set of 1060 yeast phosphopeptide sequences identified in published studies (3,4) were calculated and binned in 0.5 pI-unit intervals. The result are shown in figure 6-1a. Phosphorylation typically shifts the pI value of a peptide to a lower value due to the negative charge of the moiety. As can be seen, there are regions of pI space to which a phosphorylated peptide might be more apt to focus, a fact that was utilized by Yates' group in a study of the tumor necrosis factor pathway (5). It must be remembered, however, that the stoichiometric ratio of a phosphopeptide to its unmodified equivalent is at least 1/100, and perhaps as low as 1/1000 for the case of phosphotyrosine. Even a 100-fold enrichment would still result in an overabundance of non-phosphorylated peptide in almost all regions of pI space, limiting the use of pI as an enrichment strategy in this manner.

If the other negatively charged groups on a peptide that are not phospho-specific are eliminated, however, the differential between the phosphorylated and non-phosphorylated forms of a peptide may be expected to be much larger. Methylesterification, for example, is a reaction that neutralizes the acidic groups of a peptide, and is commonly utilized in the IMAC enrichment strategy. Figure 6-1b shows the distribution of theoretical pI values for the same group of peptides assuming that the peptides are fully methylesterified. As can be seen, all pI values are shifted to higher pI values. But the pI values of the non-phosphorylated forms of the peptides are shifted

much higher, resulting in almost all phosphopeptides located below a pI of 8.5 or 9.0. Isolating the lower pI fractions of isoelectrically focused methylesterified samples should provide a significant enrichment for phoshopeptides. Studies are ongoing to confirm the feasibility of this type of enrichment strategy. The difficulty in the experiment is getting the methylesterfication reaction to go to completion.

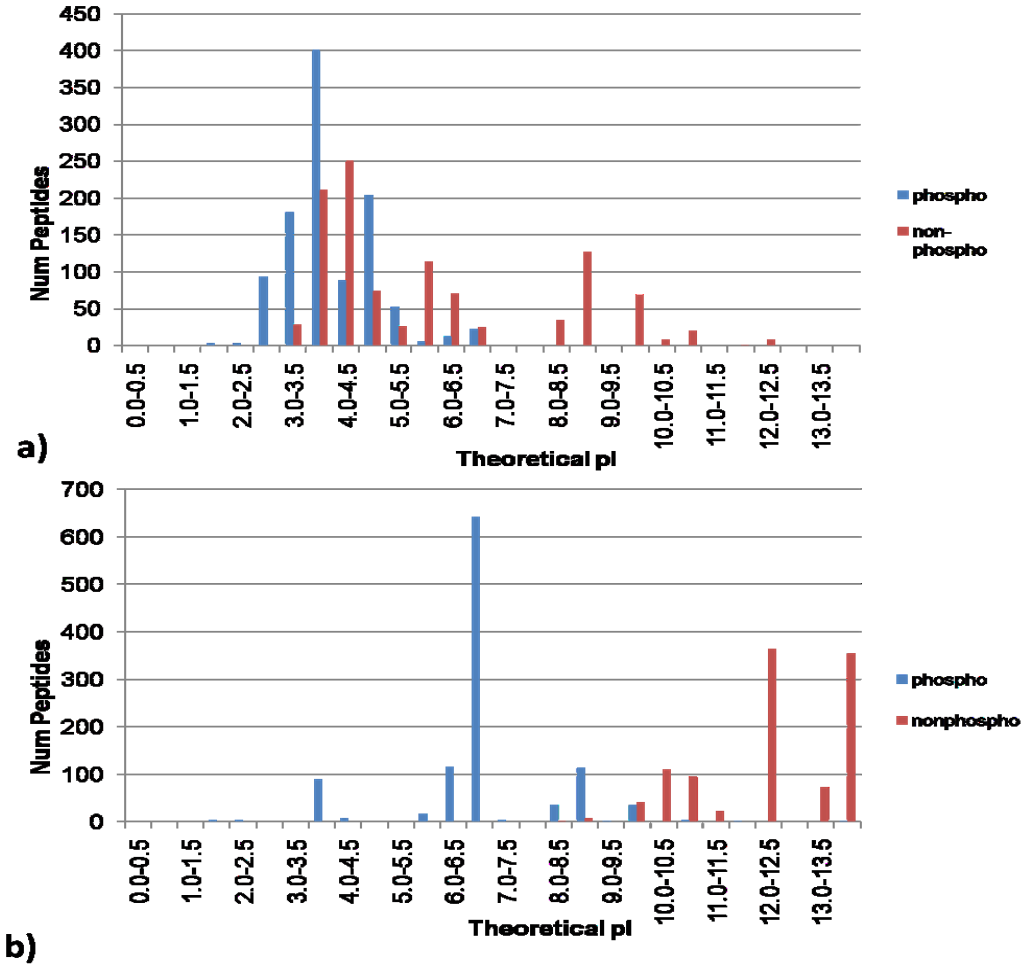


Figure 6-1. Illustration of a phosphopeptide enrichment strategy via isoelectric focusing. a) Distribution of calculated pI values for 1060 yeast peptides in their phosphorylated forms (blue) and non-phosphorylated forms (red). b) Distribution of the same peptides but assuming complete methylesterification.

6.3 Machine Learning

In Chapter Two of this thesis, several machine learning algorithms were introduced that address the problem, "can a peptide identification generated by a database search engine be classified as being correct or incorrect in an automated way with a high degree of confidence?" The answer, in short, was 'yes'. The pattern classification algorithms, given simply a table of data, were able to achieve a sensitivity and specificity matching PeptideProphet and perform significantly better than published linear thresholding methods. The caveat of those results was that the tools required a good set of training data: the algorithms are supervised approaches. Fully labeled training data are not trivial to acquire, however, and such data would likely be specific to a particular instrument type and run methodology. The PeptideProphet algorithm on the other hand is an example of an unsupervised approach: it can be run on a new dataset without the need to provide a set of training data. The algorithm is thus amenable to more routine use in a typical lab environment.

PeptideProphet, however, does rely on prior information. It is an example of a generative model in that it fits distributions to the data and calculates probabilities based on those distributions. The model works well insofar as the distributions model the data sufficiently accurately. It turns out that this is a good assumption in most situations, as PeptideProphet has been shown to successfully model many mass spectrometry data sets. The boosting and random forest machine learning algorithms, on the other hand, are examples of discriminative models: they make no assumptions about the underlying distribution of data. It is an opinion in machine learning that a discriminative approach is more general and less-risky when handling new types of data, exemplified by the statement by Vapnik (6) that "one should solve the [classification] problem directly, and never solve a more general problem as an intermediate step", the general problem discussed being the calculation of a $p(x|y)$ empirical distribution, such as the 'correct' and 'incorrect' distributions of PeptideProphet.

Of course, the PeptideProphet approach is one that is easily implemented in an unsupervised manner such as the expectation maximization (EM) method implemented by the tool. This is a less-straightforward problem using the discriminative approach, and the requirement for a global solution of this problem to be unsupervised trumps any performance degradation that may result from a less-than-optimal fit of the underlying model distributions. However, perhaps some of the difficulties associated with using the PeptideProphet algorithm that occur can be addressed by introducing some aspects of discriminative models into the framework. Moreover, if a discriminative model implementation such as logistic regression, boosting, etc. could be formulated with a sufficient unsupervised component to make it easily adaptable to various data sets and search engine outputs, a more optimal solution to the mass spectrometry database search result classification problem could be achieved.

6.4 Application of a mixture model classification strategy: predicting zymogen granule membrane topology

The mixture model approach (as used in PeptideProphet) is an example of a general classification strategy that can be useful in a variety of situations. An application of this methodology is now discussed in the context of a different problem: classifying a peptide as belonging to one of two distinct categories based on an abundance measure. This project is an illustrative example of a situation in which a simple thresholding strategy was insufficient, and an adequate solution required a more flexible and precise strategy for calculating a probability score and assigning a peptide to a state.

The overall goal of the research project is to develop an architectural model for zymogen granule membrane (ZGM) proteins. Zymogen granules (ZGs) are specialized organelles in pancreatic cells for digestive enzyme storage and regulated secretion, and are classic models for studying secretory granule function. The topological organization of a ZGM protein relative to the lipid bilayer dictates its accessibility to interacting partners and modifying enzymes. Therefore, an accurate topology model describing the

number of transmembrane spans and the orientation of a ZGM protein with respect to the membrane is essential for understanding its correct function.

The experimental methodology used for studying protein topology was a global protease protection analysis coupled with iTRAQ labeling. Two groups of isolated ZG membrane-bound proteins were compared: one group was treated with protease K and the other left untreated. The protein samples were digested with trypsin and the resulting peptides labeled with separate iTRAQ reagents. The peptides were then mixed and analyzed by 2D LC-MALDI-MS/MS. The hypothesis of the experiment was that the population of peptides from the cytoplasmic-oriented side of the ZG membrane would be exposed to protease K and would thereby be removed, resulting in a significant reduction in relative abundance when compared with their non-treated equivalents. Proteins from the luminal side of the membrane would conversely be protected from the protease and show no change in relative expression.

Plotting the identified peptides by their iTRAQ ratios does indeed show two distinct populations of peptides, one with lower ratios and one with ratios near 1.0, as expected (Figure 6-2). For these data, the bimodal distribution observed suggests that a simple thresholding heuristic could have been utilized to classify the majority of peptides as either cytoplasmic or luminal, (e.g. ratio <= 0.5 cytoplasmic). However, different replicates of the experiment indicated variability in the shapes and mean values of these distributions, defying the assignment of any specific value as a threshold. Moreover, there appeared to be significant overlap between the distributions.

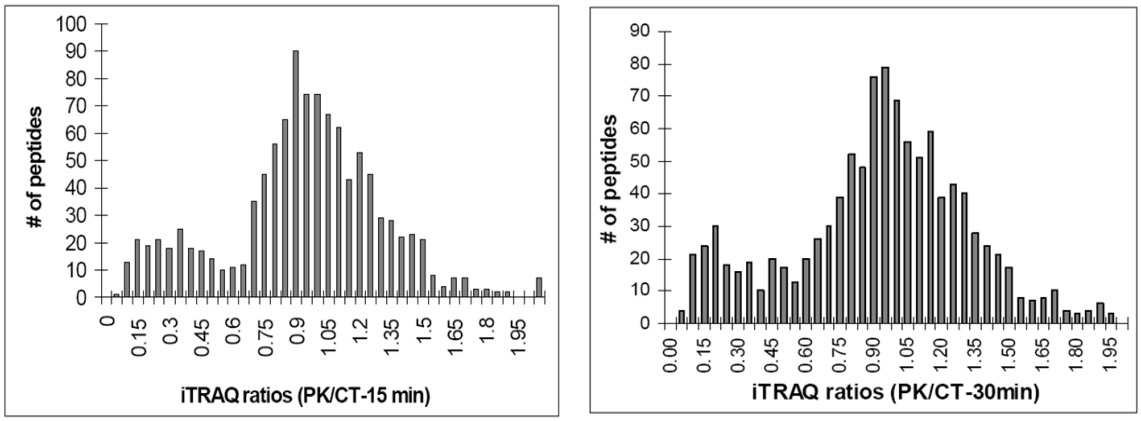


Figure 6-2. iTRAQ ratio histograms of tryptic peptides from the protease K treatment. Tryptic peptides from four different groups of ZGMs, 15 min control, 30 min control, 15 min protease K treatment and 30 min protease K treatment, were labeled with 4 different iTRAQ reagents and mixed. The histogram on the left shows 15 min protease K treatment vs 15 min control (PK/CT-15 min); the study on the right shows 30 min protease K treatment vs 30 min control (PK/CT-30 min). A total of 1079 peptides were plotted between iTRAQ ratio 0 and 1.9 with the bin size at 0.05.

Based on the assumption that peptides must be derived from one of only two categories, cytosolic or luminal, the model fits two distributions to the data set in an iterative manner. The model then estimates the probability that any observed peptide is derived from either the luminal or the cytoplasmic fraction based on the observed iTRAQ ratio. The model is thus similar to the PeptideProphet model, or a similar mixture model developed by Alexey Nesvizhskii for studying peroxisomal membrane proteins (7, 8). Figure 6-3 shows the results of fitting two separate curves to the observed, bimodal distribution of all peptide iTRAQ ratios in each dataset. For these data, a normal curve is used to model the lower-ratio cytosolic fraction, and a gamma distribution for the luminal fraction. The distributions are fitted to the data in a "semi-supervised" manner: fundamentally an EM algorithm is used to learn an optimal fit for the distributions to the dataset in an automated way starting from training distributions of peptides known to be cytosolic and luminal, but the fit is assisted by setting guiding parameters for intensity and shape based on both training data and manual curation.

Briefly speaking, the EM algorithm is an iterative, two-step optimization approach in which the parameters of each distribution are used to calculate, at each iteration, the probability of each peptide belonging to both the cytosolic and luminal fractions. These probabilities are then used to adjust the distribution shapes by weighting the contribution of each peptide ratio to each of the two distributions in a manner proportional to its probability of being in that distribution. The algorithm proceeds by successively re-calculating probabilities after each adjustment of the curve, then adjusting the curve based on these weighted probabilities, until a convergence is achieved. The probabilities of each peptide are then reported as a Bayesian probability. For the cytosolic fraction the probability of the peptide being cytosolic given a peptide ratio r is calculated as:

$$p(cyt|r) = \frac{p(r|cyt)p(cyt)}{p(r|cyt)p(cyt) + p(r|lum)p(lum)}$$

The numerator of this formula may be interpreted as the probability of having the iTRAQ ratio of r and being cytosolic, and the denominator as the overall probability of having the iTRAQ ratio r. The p(r|cyt) and p(r|lum) terms are the values calculated for the cytosolic and luminal distributions, respectively, at a given value of r; finally, the p(cyt) and p(lum) terms are the "prior" proportion of cytosolic to luminal peptides in the dataset. The luminal probability at any value of r is calculated in an analogous manner. For these data, the probabilities of peptides known with high-confidence to be either cytosolic or luminal were fixed to '1' for the corresponding fraction (and '0' to the other), and modeled in combination with the unknown peptides. This had the effect of guiding the model learning, a method utilized in Marelli et al. as well (7).

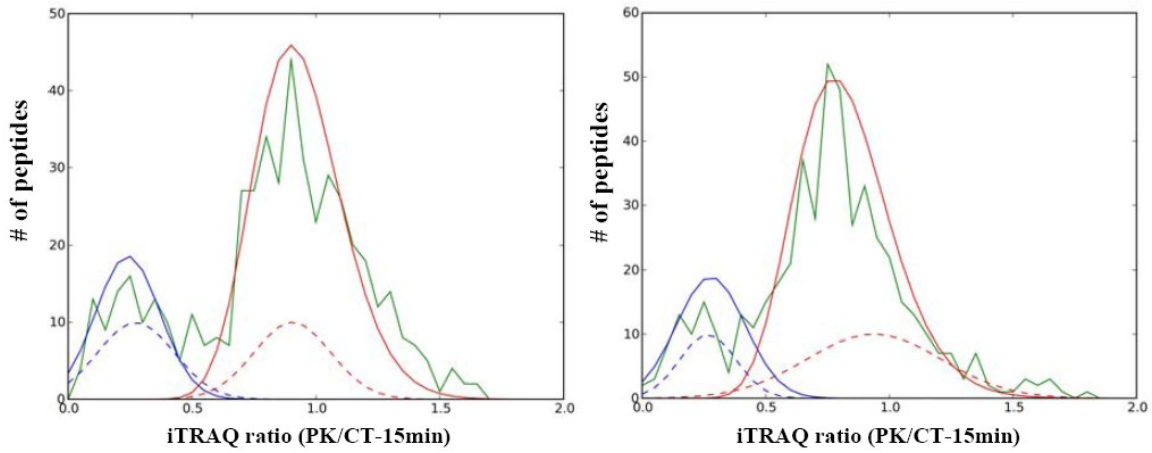


Figure 6-3. Learned model fits to iTRAQ ratio distributions. Shown are histograms for iTRAQ ratios for two independent experiments. The distribution of all peptides in each dataset is shown as green. The two final curves learned by the model for both luminal (red) and cytosolic (blue) distributions are plotted. The dashed curves indicate the histograms of the training datasets, used as a starting point for the model. The training distributions have been scaled by a factor of ten.

The statistical model thus provides a way to calculate probabilities for the assignment of a topological category for each identified peptide based on its iTRAQ ratio. The approach provides a method for deriving a global topology map of ZG membrane proteins that is flexible enough to accommodate variations in data sets due to experimental differences such as protease exposure time or changes in mass spectrometry run conditions. This model provides a foundation for developing a higher order architecture model of the ZGM and for future functional studies of each individual ZGM protein.

6.5 Phosphoproteomics

It is difficult to overestimate the importance of phosphorylation in the understanding of cellular biology and disease. The fact that an estimated 518 kinases and 130 phosphatases are encoded in the human genome underscores the prominence of the modification in cellular processes. Unregulated cell differentiation is the defining feature

of all cancers, and this differentiation occurs primarily as a result of perturbed signaling. In fact, protein tyrosine kinases (PTKs) are the largest group of dominant oncogenes with structural homology (9). It is no coincidence therefore that one of the largest groups of drug targets are protein kinases, in particular PTKs (10). The effectiveness of kinase inhibitor drugs such as imatinib (Gleevec), gefitinib (Iressa), and erlotinib (Tarceva) attest to the therapeutic relevance of kinases, and by association, phosphorylation mechanisms.

Developing global assays that can monitor phosphorylation pathways is therefore an important biological goal. Recent improvements in the ability to enrich for phosphopeptides have facilitated new avenues for analysis, based on global approaches. For example, the increase in the number of phosphorylation site identifications has permitted evaluation of global phosphorylation consensus sequence patterns (11-14). These sequence motifs are important in that they may be diagnostic of the kinase responsible for modifying the protein, implicating the identified protein as having been targeted by a specific kinase (10, 11). Such information is useful for mapping signaling pathways and understanding the functional role of the protein. The identification of specific phosphorylations on kinases themselves can provide valuable diagnostic information, in that the site of phosphorylation can specify specific states of activation of that kinase. MAP kinases (MAPKs), for example, are activated by phosphorylation of two sites within an "activation loop" of the kinase, at a threonine and a tyrosine residue with the motif TEY (rat Erk2, yeast Slt2), TPY (stress activated protein kinases), or TGY (p38, Hog1), or TNY (yeast Smk1) (15). These domains have been identified in proteomics studies (16); the fact that the MAPKs were in activated states when detected could be of significant biological interest in a study. A global phosphoproteomics study of drug treated vs control mice bearing human A431 tumors expressing high levels of the EGF receptor showed significant modulation of the EGFR signaling pathway. The study found 50 phosphopeptides differentially expressed as a function of treatment with the EGFR inhibitor, and a new tyrosine phosphoylation site on EGFR itself (17). Mann et al. studied the EGFR pathway as well, focusing specifically on tyrosine phosphopeptides , and identified five new phospho-Y sites on different members of the pathway (18). In a

later study, the group used SILAC to study temporal changes in the EGFR pathway, identifying virtually all known epidermal growth factor receptor substrates and 31 novel pathway components, as well as the time course of their activation upon epidermal growth factor stimulation (19). In all, the necessity to understand the details of specific, localized phosphorylation events is critical for understanding global mechanisms of signaling and cellular control. Global phosphorylation studies are able to provide increasingly rich datasets that provide significant insights into these mechanisms, with promising clinical therapeutic usefulness.

There are still significant methodological and informatics challenges to large-scale phosphoproteomic studies, however. The intent of a large portion of this thesis was to develop methodologies for addressing several of these issues. One major issue is one of sensitivity. Given that phosphopeptides are frequently of lower-abundance in the cell, their detection is correspondingly more difficult. This is an issue that is more poignant for phospho-tyrosine (pY), given that it is an order of magnitude rarer than serine and threonine phosphorylation. Enrichment strategies have transformed our ability to study phosphorylated protein and peptide forms, however, and it is hoped that these approaches will soon be efficient enough to permit phosphorylation studies on samples obtained in clinical studies. This goal can be facilitated by the development of computational strategies that optimize the use of the data that is generated.

Data-dependent neutral loss scanning addresses the issue of poor phosphopeptide fragmentation in an instrument, and remains a popular technique. The work here described a new and effective computational strategy for analyzing these data sets, demonstrating an increase in phosphopeptide identification sensitivity. The mass spectrometry methodologies were investigated as well, verifying the effectiveness of MSA as an instrument technique for generating data, and comparing data processing methodologies for handling data generated by both this and data-dependent MS^3 datasets. Datasets of these and various other instrumentation approaches will continue to be generated, and the ability to mine them effectively is consequently a significant one.

As new methodological work is developed, there will continue to be a need for informatics strategies that complement specific data generation techniques. For example, the problem of very low phosphotyrosine abundance is often enhanced in large scale studies in that the pY moiety is more stable than the pS or pT and thus resistant to neutral loss (18). A scanning approach in which the pY immonium ion (216.043 Da) is detected and used to direct the acquisition of MS/MS spectra has proven a useful strategy, dramatically increasing sensitivity (17). The capability to do this type of scanning on an LTQ instrument using the new HCD methodology may have a significant impact. An informatics approach which accommodates this type of information in peptide scoring might be useful. Other instrument approaches may prove relevant as well. It has been suggested by Wolf Lehmann that in the case of CID, the poor backbone fragmentation that results upon loss of the pSer/pThr phosphate moiety is a function of a low collision offset setting; increasing the collision offset induces multistep fragmentations which increases backbone fragmentation (19). Instrument settings which are tuned to generate particular spectra characteristics can be assessed using the techniques described in this work, with the goal of better optimization.

As the data from large scale studies become more available, informatics techniques which mine data sets for detection of specific target moieties such as the kinase activation signatures mentioned above are another possibility. These types of data demonstrate the need to have efficient, automated methods for not only identifying phosphopeptides, but confidently localizing the sites of phosphorylation. Phosphorylation site localization was a specific theme in this thesis. An implemention and of a published methodology for calculating a confidence score in the site localization for all phosphopeptides identified in a study was described, including an extension of the work to enable this functionality for MSA and MS^3 spectra. The method was used in the comparison of the overall mass spectrometry methodologies, but is suitable for use as a site-localization scoring implementation in any phosphoproteomics study.

Other aspects of large-scale phosphoproteomics data analysis have not been sufficiently explored. There are a number of well-defined sources for false positive identifications in MS^2/MS^3 data (19); the degree to which these errors occur in large data

sets has not been sufficiently examined. Examples of these are 'close to 98/z' losses of proline (97.053/z), valine (99.068/z), threonine (101.048/z), or cysteine (103.009/z) that could easily be mistaken for a phosphate loss. Neutral loss of these amino acid residues occur, particularly on the N-term in the presence of a nearby proline as discussed in Chapter 1 (see also Supplemental Table ST-1 in the Appendix). Methionine oxidation is a source for false positive phosphopeptide identifications as well. The loss associated with this PTM is methanesulfenic acid of mass 64/z, and is highly efficient (20). The doubly-charged loss (64/2) is easily mistaken for the triply-charged (98/3) phosphoric acid in a data-dependent methodology. In this event, the charge state of the neutral loss peak does not correspond with the charge state of the peptide precursor, and could easily be detected with diligent data analysis. However, the study in (19) provides an example in which MS^3 spectra searched with the -18 Da dehydroalanine or dehydrobutyric acid Ser or Thr residue modifications produced significantly scoring false positive identifications. Only when a search was conducted using a variable modification of -48 Da on the Met residue did the correct peptide get assigned to the spectra. This type of error is more difficult to interpret when either MS^2 or MS^3 data alone are interpreted, however, and provide an argument for having both types present.

6.6 Final Remarks

In the late 1980's, mass spectrometry of proteins was a novel technique in biological research. Now, twenty years later, it has matured to the point that it is the standard technique for identifying and quantifying proteins in high-throughput experiments. The very rapid pace of progress in this field has resulting in an explosion of data, which continues to be generated at an ever-increasing rate. An entire field of research has resulted, based on the need for automated computation to make use of the mountain of information.

In one sense, the main body of work in the proteome informatics domain was motivate simply by the necessity to keep pace with the data being produced, providing a

useable means to interpret it. Once the fundamental search framework was in place, refining it became the most significant task with automation improvements and scoring improvements as the focus of the bulk of informatics research in this area. As new methodologies were introduced such as quantification information, the software accommodated these data in response. Any changes to the established paradigm can produce difficulties, however. A SEQUEST- or Mascot-equivalent algorithm for top-down analysis, or a search engine which has been optimized for different ionization techniques such as ECD or ETD, has yet to be established. Cross-linking studies can confound standard search approaches, and many search tools cannot effectively reward very high mass accuracy data. Although progress in mass spectrometry data interpretation is very significant indeed, there is a large need for continued development.

Directed enrichment studies are becoming more established, with large-scale efforts that target specific regions of protein space, such as the 'phosphoproteome' or the 'glycoproteome' of various organisms, organelles, tissues or samples. The need for focused computational efforts which are correspondingly directed at these particular sub-domains, accommodating the rules and nuances that arise that are specific to these data, are necessary. The work in this thesis begins to address this issue, presenting informatics strategies directed at newer data types and methods in the phosphoproteomics domain. It is very likely that, rather than the gross 'highly expressed or suppressed' abundance measurement of a protein, it is in the more subtle modification state of a protein, or in a more complicated modification 'pattern' in a set of proteins, where the most significant biological insights will occur. When the computational strategies for interpretation of these types of data become more sensitive, the likelihood of detecting these more subtle levels of control will be possible.

To be most useful, proteomics data should not be interpreted in isolation. An enormous amount of complementary information can be derived from genomics and metabolomics analyses of a biological system. The ultimate goal of any of the 'omics' fields is to provide the fundamental information necessary to understand a biological mechanism. What is most necessary for this higher-order understanding is a functional representation of the system being studied, a framework which describes the connection

between the fundamental components being measured in a study—the genes and the proteins—to the higher level organizational, functional states, networks and pathways that dictate the phenotype. Network modeling approaches are increasingly successful in elucidating the fundamental motifs which control biological systems. Data sets in which gene expression, protein expression and metabolite expression have been measured for a common system are now being generated, and should provide a more holistic view of biological problems.

In conclusion, we are at the brink of being able to develop a newer, higher-order understanding of biological systems. We are just now developing the ability to assemble the complex, diverse information that proteomics and other high-throughput quantitative technologies are generating into systems-level conceptual frameworks that are biologically relevant, accurate, and predictive. The end result will be the ability to model a "Virtual Cell", with the power to expose the biological complexity of a system in a more usable and comprehensible way. The work presented in this thesis will facilitate the translation of protein expression data into higher-order biological knowledge.

References

1. Bradshaw, R. A., Burlingame, A. L., Carr, S., Aebersold, R. (2006) Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics*. 5: 787-8

2. Malmström J, Lee H, Nesvizhskii AI, Shteynberg D, Mohanty S, Brunner E, Ye M, Weber G, Eckerskorn C, Aebersold R. (2006) Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J Proteome Res.* 5: 2241-9.

3. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 4, 310-327.

4. Ballif, B. A., Villén, J., Beausoleil, S. A., Schwartz, D., Gygi, S. P. (2004) Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* 3: 1093–1101.

5. Cantin, G. T., Venable, J. D., Cociorva, D., Yates, J. R. III (2005) Quantitative Phosphoproteomic Analysis of the Tumor Necrosis Factor Pathway *J. Proteome Res.* 5: 127 -134.

6. Vapnik, V. N. (1998) *Statistical Learning Theory*, John Wiley & Sons.

7. Marelli, M., Smith, J. J., Jung, S., Yi, E., Nesvizhskii, A. I., Christmas, R. H., Saleem, R. A., Tam, Y. Y., Fagarasanu, A., Goodlett, D. R., Aebersold, R., Rachubinski, R. A., and Aitchison, J. D. (2004) Quantitative mass spectrometry reveals a role for the GTPase Rho1p in actin organization on the peroxisome membrane. *J Cell Biol* 167: 1099-1112.

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

9. Blume-Jensen, P., Hunter, T. (2001) Oncogenic kinase signaling. *Nature* 411: 355-65.

10. Moran, M. F., Tong, J., Taylor, P., Ewing, R. M. (2006) Emerging applications for phospho-proteomics in cancer molecular therapeutics. *Biochim Biophys Acta*. 1766: 230-41.

11. Obenauer, J. C., Cantley, L. C., Yaffe, M. B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*. 31: 3635-41.

12. Schwartz, D. Gygi, S. P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol.* 23: 1391-98.

13. Villén, J., Beausoleil, S. A., Gerber, S. A., Gygi, S. P. (2007) *Proc. Natl. Acad. Sci. U.S.A.* 104: 1488–93.

14. Ballif, B. A., Carey, G. R., Sunyaev, S. R., Gygi, S. P. (2008) Large-Scale Identification and Evolution Indexing of Tyrosine Phosphorylation Sites from Murine Brain. *J. Proteome Res*. 7: 311-318

15. Ferrell, J. E. Jr., Bhatt, R. (1997) Mechanistic studies of the dual phoshorylation of mitogen-activated protein kinase. J. Biol. Chem 272: 19008-19016.

16. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., White, F. M. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat Biotechnol*. 20: 301-5.

17. Stover, D. R, Caldwell, J., Marto, J., Root, K., Mestan, J., Stumm, M., Ornatsky, O., Orsi, C., Radosevic, N., Liao, L., Fabbro, D., Moran, M. F. (2004) Differential phosphoprofiles of EGF and EGFR kinase inhibitor-treated human tumor cells and mouse xenografts, *Clin. Proteomics* 1: 69–80.

18. Blagoev, B., Ong, S. E., Kratchmarova, I., Mann, M. (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol*. 22: 1139-45.

19. Lehmann, W. D., Krüger, R., Salek, M., Hung, C. W., Wolschin, F., Weckwerth, W. (2007) Neutral loss-based phosphopeptide recognition: a collection of caveats. *J Proteome Res*. 6: 2866-73.

20. Morgenthal, K., Weckwerth, W. (2006) Methionine oxidation in peptides--a source for false positive phosphopeptide identification in neutral loss driven MS^3. *Rapid Commun Mass Spectrom*. 20: 2516-8.

Appendix

Appendix

Supplemental Table ST-1. Unique instances of N-term amino acid neutral loss in the 9-Mix dataset from Chapter 2.

ms2-sequence	ms3-sequence	charge	#occurrences
K.DSADGFLK.I	D.SADGFLK.I	1	3
L.PDPQESIQR.A	D.PQESIQR.A	2	14
Y.APELLYYANK.Y	A.PELLYYANK.Y	2	4
K.TVM[147]ENFVAFVDK.C	V.M[147]ENFVAFVDK.C	2	2
K.VVAGVANALAHR.Y	V.AGVANALAHR.Y	2	9
K.DNPQTHYYAVAVVK.K	S.PQTHYYAVAVVK.K	2	24
K.YICDNQDTISSK.L	I.CDNQDTISSK.L	2	25
K.TVMENFVAFVDK.C	V.MENFVAFVDK.C	2	8
K.ELPDPQESIQR.A	L.PDPQESIQR.A	2	53
F.YAPELLYYANK.Y	A.PELLYYANK.Y	2	31
K.LKPDPNTLCDEFK.A	K.PDPNTLCDEFK.A	2	11
R.ETYGDMADCCEK.Q	T.YGDMADCCEK.Q	2	1
K.DDPHACYSTVFDK.L	D.PHACYSTVFDK.L	2	17
N.IPMGLLYSK.I	I.PMGLLYSK.I	2	17
K.VLVLDTDYKK.Y	L.VLDTDYKK.Y	2	2
C.VPCADQSSFPK.L	V.PCADQSSFPK.L	2	5
K.YIPIQYVLSR.Y	I.PIQYVLSR.Y	2	2
R.MPCTEDYLSLILNR.L	M.PCTEDYLSLILNR.L	2	2
R.TPEVDDEALEKFDK.A	T.PEVDDEALEKFDK.A	2	3
R.TPEVDDEALEK.F	T.PEVDDEALEK.F	2	12
R.KVPQVSTPTLVEVSR.S	V.PQVSTPTLVEVSR.S	2	7
S.CVPCADQSSFPK.L	V.PCADQSSFPK.L	2	6
K.EFTPVLQADFQK.V	T.PVLQADFQK.V	2	7
R.IIPGFMCQGGDFTR.H	I.PGFMCQGGDFTR.H	2	8
R.ETYGDM[147]ADCCEK.Q	T.YGDM[147]ADCCEK.Q	2	3
K.DNPQTHYYAVAVVK.K	D.NPQTHYYAVAVVK.K	2	7

K.SVDDYQECYLAM[147]V.P	V.DDYQECYLAMVPSHAVVAR.	2	5
T.DAENCHLAR.G	D.AENCHLAR.G	2	4
L.AM[147]AASDISLLDAQSAPL	A.M[147]AASDISLLDAQSAPLR.	2	1
R.LACGVIGIAK.-	A.CGVIGIAK.-	2	2
K.TSHMDCIK.A	S.HMDCIK.A	2	4
L.LEACTFHKP.-	E.ACTFHKP.-	2	2
K.DGPLTGTYR.L	G.PLTGTYR.L	2	1
K.EDVIWELLNHAQEHFGK.D	D.VIWELLNHAQEHFGK.D	2	2
K.ELPDPQESIQR.A	D.PQESIQR.A	2	1
K.SVTDCTSNFCLFQSNSK.D	V.TDCTSNFCLFQSNSK.D	2	1
R.TVGGKEDVIWELLNHAQEHF	V.GGKEDVIWELLNHAQEHFGK	3	14
R.TPEVDDEALEKFDK.A	T.PEVDDEALEKFDK.A	3	9
T.YFPHFDLSHGSAQVK.G	F.PHFDLSHGSAQVK.G	3	1
D.NPQTHYYAVAVVK.K	S.PQTHYYAVAVVK.K	3	2
K.WSGPLSLQEVDERPQHPLQ	S.GPLSLQEVDERPQHPLQVK.	3	2
K.DAIPENLPPLTADFAEDKDVC	I.PENLPPLTADFAEDKDVCK.N	3	2
V.DDYQECYLAMVPSHAVVAR.	Q.ECYLAMVPSHAVVAR.T	3	1
K.DNPQTHYYAVAVVK.K	S.PQTHYYAVAVVK.K	3	1
K.AVEHLDDLPGALSELSDLHA	E.HLDDLPGALSELSDLHAHK.L	3	1