A PHOSPHOPROTEOMIC STUDY OF INSULIN SIGNALING PATHWAY USING A NOVEL HIGH-THROUGHPUT PIPELINE

By

Kebing Yu

B. Sc, Peking University, Beijing, China 2000

A Dissertation Submitted in Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy

in the Department of Chemistry at Brown University

Providence, Rhode Island

May 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This dissertation by Kebing Yu is accepted in its present form by

the Department of Chemistry as satisfying the dissertation requirements for

the Degree of Doctor of Philosophy

Date _____ _____

Dr. Arthur R. Salomon, Director

Recommend to the Graduate Council

Date _____ _____

Dr. Shouheng Sun, Reader

Date _____ _____

Dr. Carthene R. Bazemore-Walker, Reader

Approved by the Graduate Council

Date _____ _____

Dean of the Graduate School

Curriculum Vitae

Kebing Yu was born in Ningbo, Zhejiang Province of China in 1982. After completing his high school study in Ningbo, he moved to Beijing and attended Peking University as an undergraduate student in 2000. He spent four years in the College of Chemistry and Molecular Engineering (CCME) to study chemistry. In 2004, he received Scientific Innovation Prize and Bachelor of Science degree from Peking University with the thesis: *Influence of Divalent Metal Ions on the Aggregation of Mixed Egg Yolk Phosphatidylcholine*. After graduation, he continued to Providence, RI and started his Chemistry Ph.D program at Brown University. He joined Dr. Salomon's group and conducted research in the field of proteomics. He was interested in developing LC/MS based analytical methods to study protein phosphorylation in cell signaling pathways, as well as creating bioinformatic tools to analyze large-scale proteomic data.

Publications:

1) **K. Yu**, A. Salomon (2009). "HTAPP: High-throughput autonomous proteomic pipeline." *Proteomics.* **In review**
2) **K. Yu**, A. Salomon (2009). "PeptideDepot: Flexible relational database for visual analysis of quantitative proteomic data and integration of existing protein information." *Proteomics.* **In press**
3) R. Jianu, **K. Yu**, L. Cao, V. Nguyen, A. Salomon, D. Laidlaw (2009). "Effective visual integration of quantitative proteomic data, pathways, and protein information." *Trans. on Vis. and Comp. Graphics.* **In review**
4) **K. Yu**, A. Sabelli, L. DeKeukelaere, R. Park, S. Sindi, C. Gatsonis, A. Salomon (2009). "Integrated platform for high-throughput statistical and manual validation of tandem mass spectra." *Proteomics.* **9**(11): 3115 - 3125

5) V. Nguyen, L. Cao, J. T. Lin, N. Hung, A. Rit, **K. Yu,** R. Jianu, B.J. Raphael, S. Ulin, D.H. Laidlaw, L. Brossay, A. Salomon (2009). "A new approach for quantitative phosphoproteomic dissection of signaling pathways applied to T cell receptor activation." *Mol. Cell. Proteomics.* **In press**

6) I. Lee, A. Salomon, **K. Yu**, L. Samavati, P. Pecina, A. Pecinova, M. Huttemann (2009). "Isolation of regulatory-competent, phosphorylated cytochrome c oxidase." *Methods Enzymol.* **457**: 193-210.

7) H. Yu, I. Lee, A. Salomon, **K. Yu**, M. Huttemann (2008). "Mammalian liver cytochrome c is tyrosine-48 phosphorylated in vivo, inhibiting mitochondrial respiration." *BBA-Bioenergetics*, **1777**(7-8): 1066-71.

8) L. Cao, **K. Yu**, V. Nguyen, Y. Kawakami, T. Kawakami, A. Salomon (2007). "Quantitative Phosphoproteomic Analysis of Mast Cell Signaling." *J. Immunology*, **179**(9): 5864-5876.

9) T. Nuhse, **K. Yu**, A. Salomon (2007). "Isolation of Phosphopeptides by Immobilized Metal Affinity Chromatography." *In Cur. Prot. Mol. Biol.*, (Ausubel et al., eds.) 18.13.1-18.13.23. John Wiley & Sons, Hoboken, N.J.

10) I. Lee, A. Salomon**, K. Yu**, J. Doan, L. Grossman, M. Huttemann (2006). "New Prospects for an Old Enzyme: Mammalian Cytochrome c Is Tyrosine Phosphoryalted In Vivo." *Biochemistry.* **45**(30): 9121-9128.

11) **K. Yu**, Z. Yang, L. Zhang, F. Wang, S. Wen, J. Wu (2003). "The Influence Of Divalent Metal Ions On The Aggregation Properties Of EYPC." *Acta Phys.-Chim. Sin*, **19**(8): 747.

Abstract

Recent advances in the speed and sensitivity of mass spectrometers and analytical methods, the exponential acceleration of computer processing speeds, and the availability of genomic databases from an array of species and protein information databases have led to a deluge of proteomic data. Unfortunately, this enhancement in data acquisition has not been accompanied by a concomitant increase in the availability of tools allowing users to rapidly assimilate, explore, and analyze this data and adapt to a variety of experimental workflows with minimal user intervention. Often the manual aggregation and analysis of proteomic data in current proteomics software distract investigators from the biological meaning of their data, leading to the all-too-frequent deposition of proteomic data into the scientific literature with little or no biological or clinical interpretation.

We seek to fill the gap by providing a flexible platform for high-throughput autonomous proteomic analysis with the following critical components: liquid chromatography/mass spectrometry (LC/MS) acquisition control module, tandem mass spectra (MS/MS) database search, peptide spectral validation, peptide quantitation, quantitative data exploration tool within a relational database, cached public protein information databases and protein network exploration tool. The LC/MS control tool integrates lab information management system (LIMS) to provide automated multi-dimensional sample analysis, as well as captures meta-data during analysis and associates them with sample preparation protocols and experiment results in a relational database. Instrument acquired raw data are streamlined through a customized proteomic pipeline for database searching followed by peptide validation. The logistic spectral score we

developed for high-throughput statistical validation of peptide sequence assignment to MSMS spectra outperforms standard tools already available in the proteomics field such as Sequest and X!Tandem to obtain the highest yield of confident peptide assignments. The logistic spectral score outperforms SEQUEST XCorr (242% more peptides identified on average) and the X!Tandem E-Value (87% more peptides identified on average) at a 1% false discovery rate estimated by decoy database approach. Peptide identifications, along with data-dependent calculation results are directed into a relational database for organization of expansive proteomic data sets, collation of proteomic data with available protein information resources, and visual comparison of multiple quantitative proteomic experiments. This platform provides flexible adaptation to diverse workflows for the unique requirements of the individual proteomics lab, enabling proteomic scientists to modify the presentation of the proteomic data, implement extra data-dependent analysis tasks, process additional input formats and control new types of instruments.

The utility of this system is illustrated through analysis of insulin signaling pathway important to liver cancers. We explored changes in phosphorylation profile quantitatively in hIRS1-transfected NIH3T3 cells in response to insulin stimulation. In a SILAC-labeled NIH3T3-hIRS1/NIH3T3-hIRS1 Y1180F timecourse, we discovered a total of 2201 phosphorylation sites at 1% false discovery rate, among which 1862 (84.6%) were on Serine, 299 (13.6%) were on Threonine and 40 (1.8%) were on Tyrosine. Using a label-free/SILAC hybrid quantitation approach, different phosphorylation patterns were identified in wild type and mutated cell lines.

Acknowledgement

Not only at this moment, but always I'm eager to express my sincere appreciation to everyone who has helped me in preparation of this thesis, at work, and in life!

I would like to thank my graduate advisor, Professor Arthur Salomon, for his excellent supervision during these five years. I really learnt a lot from him. He is a kind mentor that provides insightful guidance but gives me a large room of freedom to work the project out. I'm feeling so lucky for being part of his group.

Thanks also go to my committee members, Professor Shouheng Sun and Professor Carthene Bazemore-Walker, and Professor William Risen who has served in my 3rd year research proposal committee.

Many thanks to all the Salomon group members – Lulu Cao, Vinh Nguyen (the BBQ master), Yiyuan Ding, Jonathan Lin, John Lung, Norris Hung, Jayameenakshi Manivannan (Meena), and former members - Cindy Banh, Sam Ulin, Anthony Sabelli. I enjoyed the time being spent together with you guys.

And to many friends in my life…

Finally, I must thank my dear parents, for their understanding, endless patience and unconditional love. Without their support, I would not have any chance here writing this document.

Table of Contents

Table of Figures

List of Tables

Table of Abbreviations

Abbreviation	Definition
MALDI	Matrix-assisted laser desorption/ionization
ECD	electron capture dissociation
ETD	electron transfer dissociation
CID	collision-induced dissociation
PTM	post-translational modification
IMAC	immobolized metal affinity chromatography
DHB	dihydroxybenzoic acid
SCX	strong cation exchange
HILIC	hydrophilic interaction chromatography
HTP	high-throughput
LTP	low-throughput
ICAT	Isotope-Coded Affinity Tags
GIST	Global Internal Standard Technology
ICPL	Isotope-Coded Protein Label
iTRAQ	isobaric tag for relative and absolute quantitation
SILAC	Stable Isotope Labeling with Amino acids in Cell culture
LIMS	Laboratory Information Management System
FDR	false discovery rate
HTAPP	High-Throughput Autonomous Proteomic Pipeline
MudPIT	Multidimensional Protein Identification Technology
TPP	Trans-Proteomic Pipeline
HPLC	High performance liquid chromatography
LTQ-FTICR	Linear Ion Trap–Fourier Transform mass spectrometer
LTQ	Linear Ion Trap
SIC	Selected ion chromatogram
BSA	Bovine Serum Albumin
CV	Coefficient of Variation
AUC	Area under curve
ROC	Receiver operating characteristic
HCC	Hepatocellular Carcinoma
IR	Insulin receptor
IRS	Insulin receptor substrate
SH2	Src homology 2
PI3K	Phosphatidylinositol 3-kinase
PDE3B	Phosphodiesterase 3B
GSK-3	Glycogen synthase kinase 3
FOXO1	Forkhead box O1
FOXO3A	Forkhead box O3A
FOXO4	Forkhead box O4

MAPK	Mitogen-activated protein kinase
HCC	Hepatocellular Carcinoma
DMEM	Dulbecco's modified Eagle's medium
NIH3T3-hIRS1	NIH3T3 cells with overexpression of human IRS1 protein
NIH3T3-hIRS1 Y1180F	NIH3T3 cells with overexpression of human IRS1 protein with Y1180F mutation
PBS	phosphate buffer saline
MeCN	Acetonitrile
TFA	trifluoroacetic acid
IGF	Insulin-like growth factor

Chapter 1

INTRODUCTION

1.1 MASS SPECTROMETRY-BASED PROTEOMICS

As an analog to genomics, the term 'proteomics' was first introduced in the late 1990s [1] to describe the large-scale study of protein identifications, structures and functionalities [2-4]. Although great progress in the human genome project leads to a huge leap in understanding human beings, it is still challenging to effectively decipher and treat various diseases based on the genomic data [5]. Some critical information, such as protein-protein interaction, relationship between protein structure and function, and dynamic post-translational modification processes, is not encoded in genes and has to be investigated at the protein level. Representing a bridge between genes and physiological functions, proteins are well recognized as the key molecules to elucidate the molecular basis of a particular cellular state.

The early stage of proteomics can be dated back to 1970s when two-dimensional gel electrophoresis was used to separate and analyze protein complex [6, 7]. By loading protein mixtures on the SDS gel, proteins were first separated by molecular weight in the first dimension, and then separated by isoelectric point in the second dimension. Initial proteomic analyses were primarily relying on silver staining to visualize protein map and Edman degradation [8] to sequence proteins, until biological mass spectrometry gained its popularity in the early 1990s [9-13]. Compared to traditional methods, the mass spectrometry-based proteomic analysis approach is more efficient, sensitive and precise [14-16]. Proteins or peptides are ionized via either electrospray [17] or MALDI [18, 19] and their mass over charge are measured. Isolated peptide ions are further fragmented to the amino acid level to provide sequence information via tandem mass spectrometry process.

Two complementary approaches are developed for proteomic analysis using mass spectrometry: the top-down and bottom-up approaches. Top-down proteomics directly analyzes samples at the protein level to provide direct information about the protein molecular weight. In the following fragmentation stage, ions are fragmented at the amide bond to yield fragment pieces that contain protein sequence information, using various approaches, notably electron capture dissociation (ECD) [20] and electron transfer dissociation (ETD) [21]. However, mass spectrometers with limited scan range and resolving power provided challenges for the analysis of large molecular weight, highly charged proteins. MS/MS spectra acquired on protein fragments are typically very complicated due to the large number of possible ways to break up protein backbones, resulting in significant difficulty to determine the protein sequence. Furthermore, the wide range of protein hydrophobicity limited the choice of possible mass spectrometry-compatible buffers that are able to dissolve samples without severe loss of certain components. On the other hand, a "bottom-up" approach focuses on analyzing small peptides, *i.e.* pre-cleaved protein fragment pieces, to identify proteins and determine details of their sequence and posttranslational modifications. Commonly used proteases, such as trypsin [22], Lys-C [23], and chemical reagents, such as cyanogen bromide (CNBr) [24], are able to digest protein primary structure at specific amino acid residues, yielding a mixture of short peptides. Although cleavage of proteins increases the number of molecules that must be analyzed, the bottom-up approach provides distinct advantages over top-down approach for certain types of analysis. First of all, most digested peptides are water-soluble, and compatible with common liquid chromatography solvents. Complex peptide mixtures can be resolved by coupling liquid chromatography directly to

mass spectrometry (LC-MS). Secondly, MS/MS spectra acquired on short peptides are much easier to interpret than those generated by proteins. Sufficient evidence regarding the sequence and post-translational modification usually could be observed on a single MS/MS spectrum.

Distinctive MS/MS spectra are produced according to the selected precursor peptide sequences and the method to generate fragment ions. Theoretically, substitution of one amino acid in the peptide sequence results in half of fragment ion masses being shifted accordingly, which would generate a totally different spectrum. On the other hand, different spectra for the same peptide could be generated due to different fragmentation methods that break different chemical bond and produce unique fragment ions (Figure 1.1) based on the internal mechanisms. For instance, collision-induced dissociation (CID) [25] usually predominantly produces b and y ions, plus a few a ions [26], while ETD produces c and z ions instead [21]. A typical CID type MS/MS spectrum is illustrated in Figure 1.2.

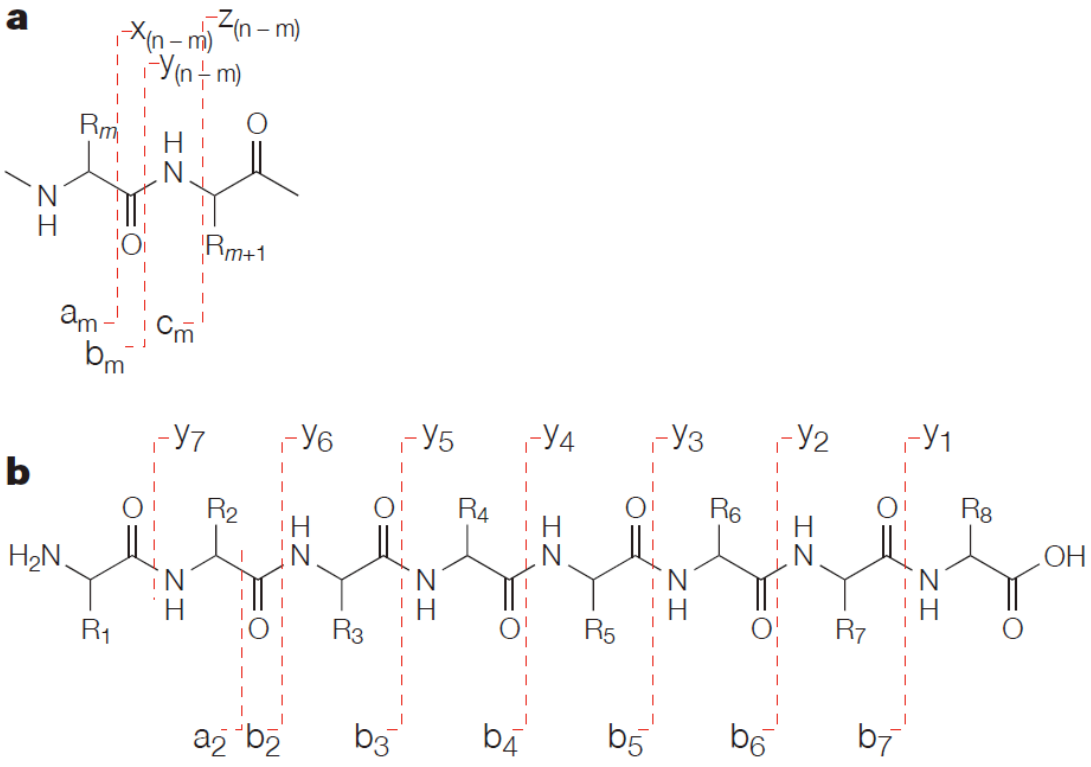


Figure 1.1: Nomenclature of peptide fragment ions. [27] a) a, b, c and x, y, z ions are defined according to the bond it breaks and the charge location. Charged N-terminal fragments are categorized as a, b, c ions and Charged C-terminal fragments are categorized as x, y, z ions. b) Ions are labeled consecutively from the original amino terminus

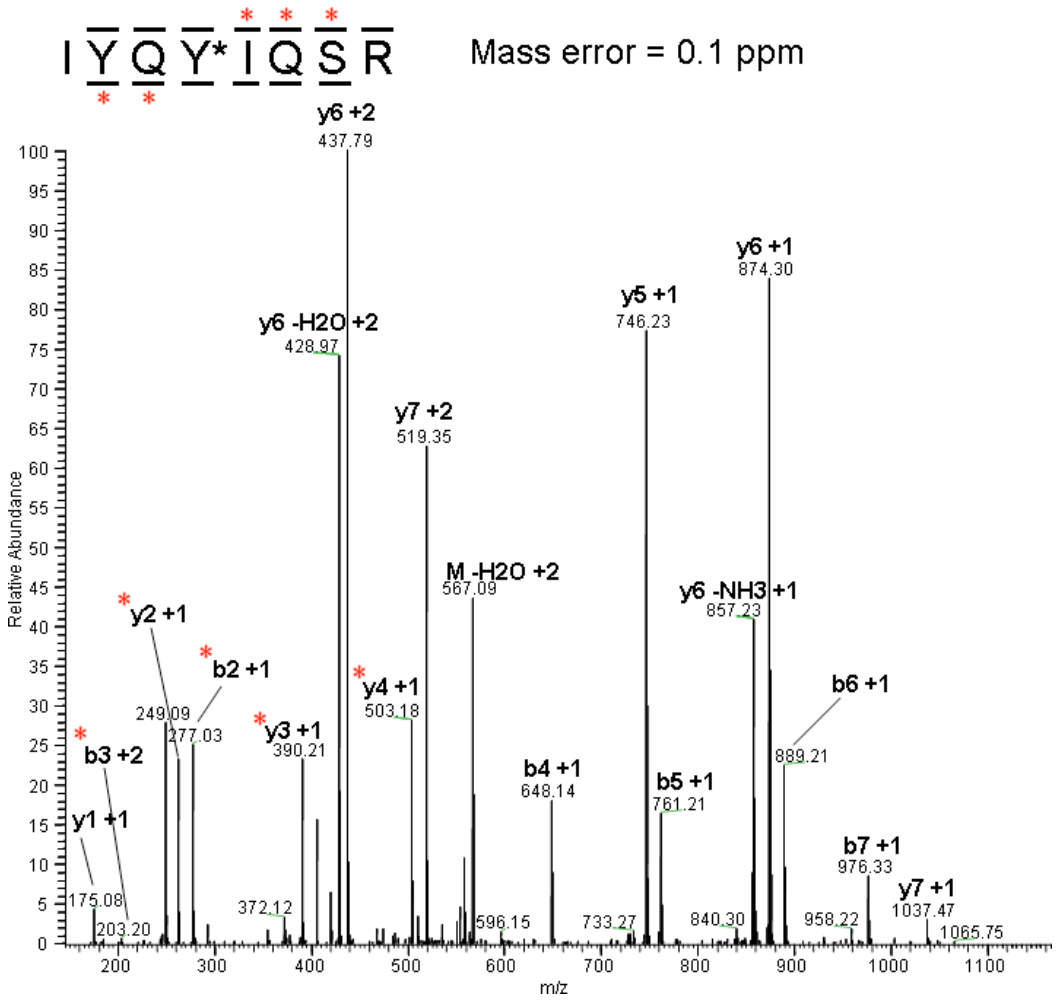


Figure 1.2: An MS/MS spectrum generated using CID fragmentation. Fragment ions are assigned to singly or doubly charged b and y ions.

Several programs are available to reconstruct peptide sequences based on MS/MS spectral information. Based on the sequencing principle, they can be divided into three classes: *de novo* [28-30], database matching [31-39], and hybrid [40-42]. Lutefisk [28] and PEAKS [29] are developed based on the *de novo* peptide sequencing concept which basically assembles the target sequence using fragment information available in the MS/MS spectra. More popular sequencing programs including Sequest [31], Mascot [33] and X!Tandem [36, 43] are based on database searching. Theoretical fingerprint spectra are generated for all possible peptides contained in a genomic protein database and

matched to the examined experimental spectra. Quality scores are computed to indicate the matching confidence between the theoretical and experimental spectra. Recently, a hybrid class of sequencing algorithm that combines *de novo* and database search is developed, represented by InsPecT [40]. With InsPecT, a short sequence tag is learned *de novo* from experimental MS/MS spectra and used to filter the broader genomic protein database to generate candidate peptides. Then a refined database search is performed on those peptides to yield the final matched sequence and modifications.

1.2 PHOSPHOPROTEOMICS

Mass spectrometry-based proteomics has been applied in many biological research areas to identify proteins, discover protein modifications and investigate protein-protein interactions. One promising work is to study cellular protein phosphorylations. Phosphorylation is one of the most important protein post-translational modifications (PTMs), which is involved in almost all cellular processes in living organisms. It is estimated that at least 30% of proteins are phosphorylated in mammalian cells, many of which are enzymatically active and regulate key physiological functions such as cell growth, apoptosis, proliferation, differentiation, and inter-/intra- cellular communication. Reversible phosphorylation results in a change in protein secondary structures, causing significant alteration of the binding affinity or activation state [44]. (Figure 1.3) Recently, many drug discovery efforts for cancers are targeted at the inhibition of enzymes called kinases and phosphatases with the ability to add or remove these modifications to

substrate proteins. Understanding the relationship between protein phosphorylation and cellular behavior and phenotype is of the first priority.

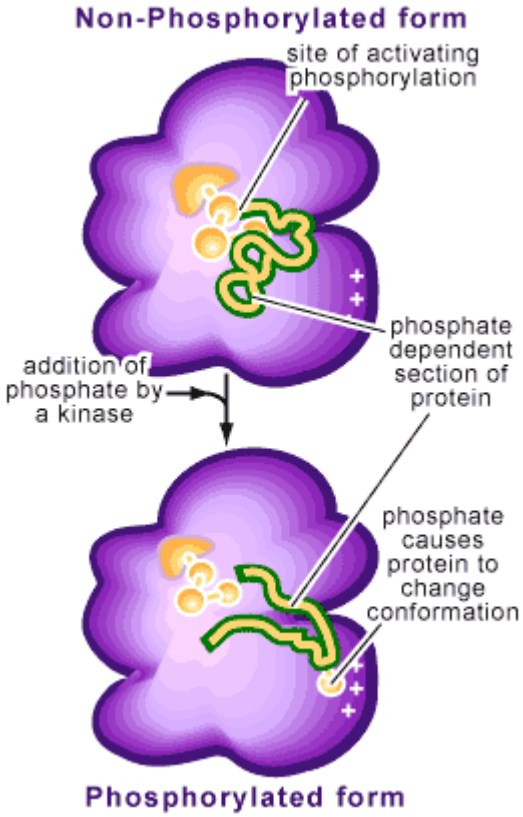


Figure 1.3: Phosphorylation induced protein conformation change. (Figure from http://www.scq.ubc.ca/?p=372)

Not too long from the first time phosphoproteomics was developed to study protein phosphorylation, milestones have been accomplished in every aspect of this field, including separation technology, mass spectrometry and bioinformatics. However, investigators still face a lot of challenges to understand protein phosphorylations thoroughly because of the lack of instrument sensitivity and method specificity in detecting phosphopeptides. This is especially true for phosphotyrosine. In normal growing cells, about 90% of phosphorylation occurs on serine, 10% on threonine and

0.05% on tyrosine [45]. Due to the low abundance of many phosphorylated proteins and substoichiometric nature of the phosphorylation of proteins, it is still not possible for the currently available instruments to resolve total phosphoproteome comprehensively. Phosphoproteomics is progressively moving forward to expand the percentage of the total phosphoproteome surveyed.

A typical phosphoproteomics analysis workflow includes several stages, among which phosphopeptide enrichment, LC/MS analysis and bioinformatics data processing are the key steps for method optimization.

As mentioned earlier, phosphopeptides are present at extremely low levels and are barely detectable without enrichment prior to MS analysis. Several enrichment strategies have been developed in the past few years. One class of enrichment methods are based on the high affinity chelation effect of phosphate group to certain metal ions, represented by Fe^{3+}-immobolized metal affinity chromatography (IMAC) [46], Ga^{3+}-IMAC [47], TiO_2 [48] and Nb_2O_5 [49]. Because of the chemically similar carboxyl functional group found on Asp and Glu residues, non-phosphopeptides having multiple acidic amino acids are competitive binders to phosphopeptides. Additional steps are required to prevent undesired contaminants in phosphopeptide enrichments. Using peptide methylation for IMAC [50] or additional organic competitor such as dihydroxybenzoic acid (DHB) for TiO_2 [48] significantly reduces non-phosphorylated peptide binding. Another enrichment strategy for phosphotyrosine is based on immunoaffinity binding between tyrosine phosphorylated protein or peptide and anti-pTyr antibody [51, 52]. Antibodies are conjugated to Protein G agarose beads and selectively capture phosphotyrosine-containing sequence from a large mixture of proteins or peptides, enabling specific

enrichment. Some phosphopeptide enrichment strategies based on chemical properties of phosphopeptides have also shown promise, such as strong cation exchange (SCX) [53] and hydrophilic interaction chromatography (HILIC) [54]. These methods enrich for phosphopeptides based on their charge state and hydrophilicity, respectively. The combination of several enrichment methods can provide enhanced selectivity in enrichment for phosphopeptides.

Improvement of sensitivity and reliability in phosphoproteomics analysis also depends on the development in mass spectrometry design. Not only because of the low abundance nature, but also due to the negatively charged phosphate group, phosphopeptides typically have poor ionization efficiency compared to non-phosphorylated peptides when the mass spectrometer is operated in positive ion mode for optimal fragmentation in MSMS experiments. Besides the sensitivity and mass accuracy, improvement in peptide fragmentation methods is equally important. O-P bond between serine/threonine residue and phosphate group is one of the most labile bonds among the phosphopeptide. With collisionally induced dissociation (fragmentation) methods (CID and CAD), energy deposited in a peptide ion is re-distributed along the whole molecule prior to backbone dissociation [25], often breaking this O-P bond and leading to neutral loss of phosphate group. (Figure 1.4 A) Dominant neutral loss peaks in MS/MS spectra significantly decreases spectral quality resulting in low signal to noise of the remaining b and y type ions. To overcome this problem, neutral-loss-triggered MS^3 [53] and pseudo-MS^3 (multi-stage activation) [55] have been developed to further fragment the neutral loss peak. (Figure 1.4 B) New fragmentation methods have been implemented as well to eliminate the neutral loss of phosphate. Phosphopeptides fragmented by ECD [56] and

ETD [21] contain less abundant neutral loss peaks because energy generated by charge neutralization is localized on a certain bond and rapidly followed by bond dissociation.

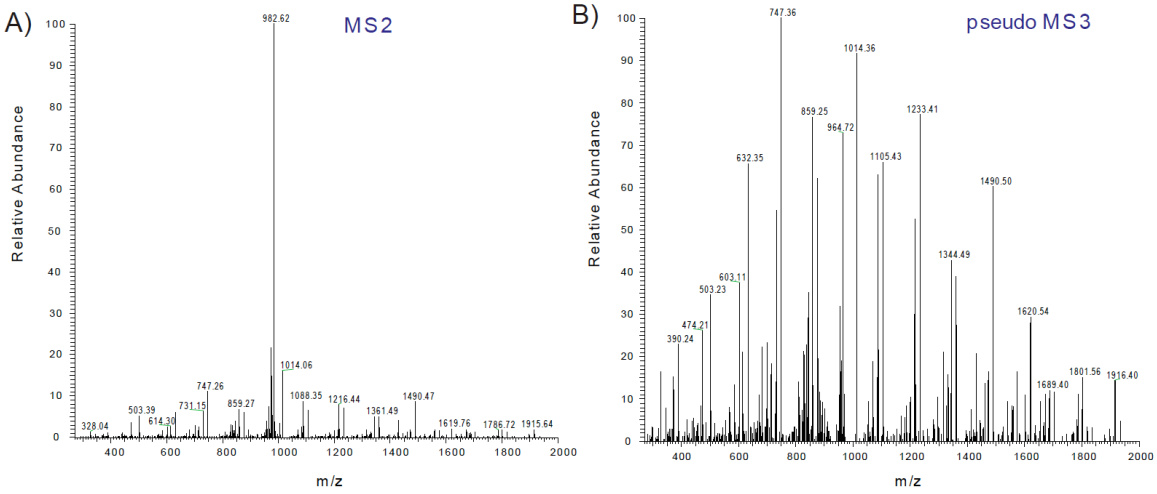


Figure 1.4: Acquired MS/MS spectra on the precursor ion m/z=1031.42. A) MS^2 only; B) pseudo MS^3. Predominant neutral loss peak (M^{2+}-OPO3, m/z: 982.62) significantly decreases the quality of MS/MS spectrum, while pseudo-MS^3 approach gives more sequence-related information.

With the ever-growing phosphoproteomic datasets, bioinformatic analysis becomes an integral part of data analysis. In a single LC/MS experiment, 10,000 MSMS spectra can be collected in a single hour of data acquisition. Many tools have been developed to improve prediction, verification and description of phosphorylation sites. Notably, Scansite searches for motifs and interaction domains that are likely to be phosphorylated based on the protein sequence motifs [57]. Search engine assigned phosphorylation sites can be validated by ascore [58, 59] in a high-throughput mode. Characterized phosphorylation sites are deposited in databases such as Phosphosite (including both high-throughput, or HTP, e.g. discovered by mass spectrometry data and low-throughput, or LTP, e.g. discovered by traditional biochemical method data) [60] and HPRD (LTP data only) [61], providing a good resource for phosphoproteomic research.

1.3 QUANTITATIVE APPROACH TO UNRAVEL SIGNALING NETWORKS

Most frequently, it is not only enough to identify the phosphopeptide and localize the phosphorylation site, but it is also critically important to resolve the temporal change of phosphorylations in response to external stimulations, or differential activation states between normal and diseased tissues. Phosphorylation and dephosphorylation events that occur in signal transduction networks are transient and dynamic, which requires a quantitative approach to capture the intermediate states. Absolute quantification is not very feasible currently, since it requires synthesis of heavy isotopic version of every peptide detected [62], which is laborious and expensive. Relative quantification is a good alternative for the purpose of quantitative comparison of several samples.

Label-free method is a cost-effective and straightforward quantitation approach [51]. To minimize experimental error, stable isotopes are also used to label different sample states and compare their intensities within a single LC/MS run. Isotopically heavy atoms can be incorporated either through chemical labeling or metabolic labeling. Chemical labeling relies on reaction of isotopic reagent with peptide functional terminals, such as cysteine, amine, or carboxylic acid groups. Peptide quantitation can be done at the MS level (ICAT [63], GIST [64], ICPL [65]), or MS/MS level (iTRAQ [66]). The advantages of the iTRAQ method is the multiplex capability (8 cellular states can be quantitatied in a single MSMS spectra) and the quantitation is performed in the MSMS scan. Peptides from different samples look identical in the MS scan and are distinguished and quantified in the MS/MS stage via a reporter group dissociated from the isobaric mass tag. Metabolic labeling with isotopically labeled amino acids is even more accurate than chemical labeling since it normalizes technical error from the very beginning of sample

preparation steps. Stable Isotope Labeling with Amino acids in Cell culture (SILAC) is a promising approach to study signaling pathways [59]. Generally, cells are cultured in special media that contains several types of amino acids that are completely replaced by their heavy isotopic versions. Lys and Arg are usually substituted because trypsin is used to digest protein in most bottom-up style proteomics experiments. Both SILAC and iTRAQ are widely used, whereas it depends on experiment condition, sample type, instrument, and budget to choose which is preferable.

It is worth mentioning that when choosing isotopic reagents, ^{15}N, ^{14}C and ^{18}O are more favorable than ^{2}H. This is because of the deuterium effect that leads to chromatography shift in retention time [67]. For relative quantitation of peptide abundance between different cellular states, co-elution of stable isotope labeled peptides provides the most accurate results and simplifies computational analysis.

1.4 OVERVIEW OF OUR FINDINGS

The primary goal of this thesis project is to provide a high-throughput solution for phosphoproteomic analysis as illustrated in Figure 1.5. To reach this goal, We have developed a flexible platform for high-throughput autonomous proteomic analysis with the following critical components: LC/MS acquisition control tool (Chapter 2), MS/MS database search (Chapter 2), peptide spectral validation (Chapter 4), peptide quantitation (Chapter 3), data exploration tool within a relational database (Chapter 3), cached public protein information databases (Chapter 3) and protein network exploration tool. (Summarized in Figure 1.6) The LC/MS control tool integrates lab information

management system (LIMS) to provide unmonitored multi-dimensional sample analysis, as well as capture meta-data during analysis to be associated with sample preparation protocols and experimental results in a relational database. Instrument acquired raw data are automatically assembled into a customized interface for database searching followed by peptide validation. The logistic spectral score we developed for high-throughput statistical validation outperforms both XCorr (242% more peptides identified on average) and the X!Tandem E-Value (87% more peptides identified on average) at a 1% false discovery rate (FDR) estimated by decoy database approach [68]. Peptide identifications, along with data-dependent calculation results are directed into a relational FileMaker/MySQL database for organization of expansive proteomic data sets, collation of proteomic data with available protein information resources, and visual comparison of multiple quantitative proteomic experiments. This platform provides flexible adaptation to diverse workflows for the unique requirements of the individual proteomics lab, enabling proteomic end-users to modify the presentation of the proteomic data, implement extra data-dependent analysis tasks, process additional input formats and control new types of instruments. The ultimate purpose of this system is to allow users to focus on extraction of biological meaning from vast data sets instead of routine data manipulation tasks.

We implement these new bioinformatic tools in the analysis of insulin signaling pathway in hepatocellular carcinoma. Using a hybrid quantitation approach combining label-free and SILAC, we were able to quantify a total of 2201 phosphorylation sites at 1% false discovery rate. Several groups of phosphorylation sites were identified based on the quantitative data processed through the high-throughput proteomic pipeline.

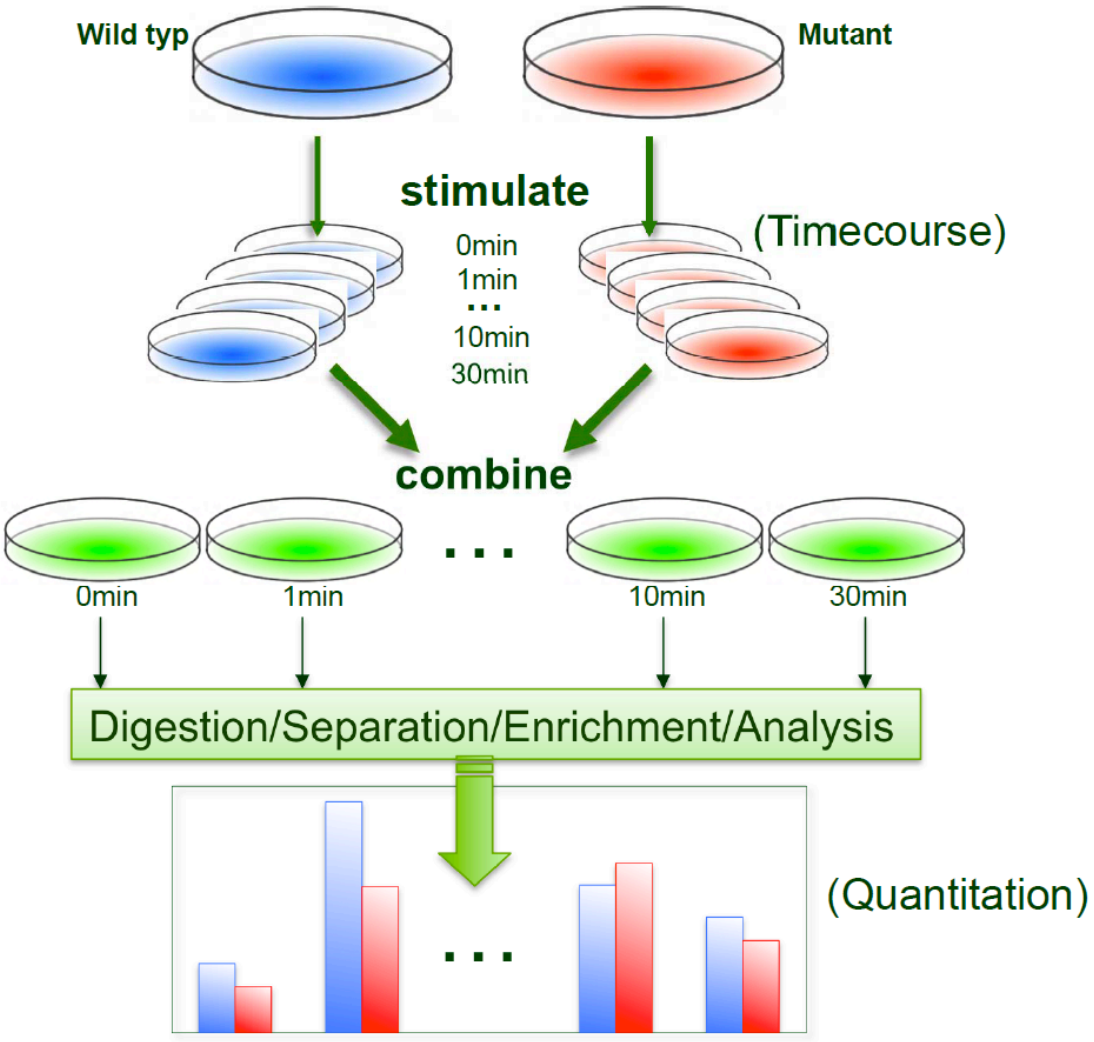


Figure 1.5: A hybrid approach in combination of label-free and SILAC quantitation methods to study phosphoproteomics in signaling pathway. Two cell lines, such as wild type and mutant, or normal and diseased are cultured in light media and heavy media separately. Then, these cells are stimulated for different time length and combined light and heavy SILAC labeled samples in 1:1 ratio to generate a SILAC timecourse.

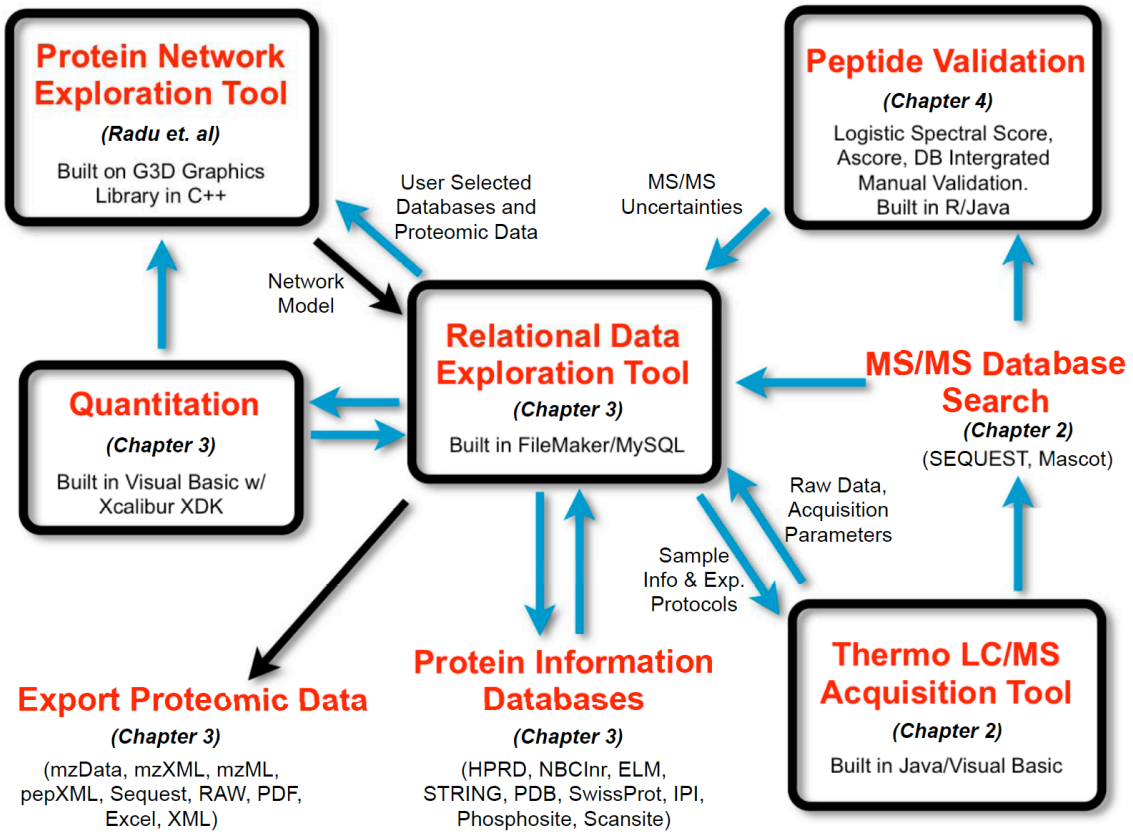


Figure 1.6: Overview of the high-throughput autonomous proteomic pipeline for phosphoproteomics. Each piece of the pipeline is described in greater detail in the indicated thesis chapter.

1.5 REFERENCE

1. James, P., Protein identification in the post-genome era: the rapid rise of proteomics. Q Rev Biophys, 1997. **30**(4): p. 279-331.

2. Anderson, N.L. and N.G. Anderson, Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis, 1998. **19**(11): p. 1853-61.

3. Liotta, L.A. and E.F. Petricoin, Beyond the genome to tissue proteomics. Breast Cancer Res, 2000. **2**(1): p. 13-4.

4. Wang, C.C. and C.L. Tsou, Post-genome Study----Proteomics. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai), 1998. **30**(6): p. 533-539.

5. Dove, A., Proteomics: translating genomics into products? Nat Biotechnol, 1999. **17**(3): p. 233-6.

6. Knopf, U.C., et al., A new two-dimensional gel electrophoresis system for the analysis of complex protein mixtures: application to the ribosome of E. coli. Mol Biol Rep, 1975. **2**(1): p. 35-40.

7. Ames, G.F. and K. Nikaido, Two-dimensional gel electrophoresis of membrane proteins. Biochemistry, 1976. **15**(3): p. 616-23.

8. Edman, P., A method for the determination of amino acid sequence in peptides. Arch Biochem, 1949. **22**(3): p. 475.

9. Chowdhury, S.K., V. Katta, and B.T. Chait, Electrospray ionization mass spectrometric peptide mapping: a rapid, sensitive technique for protein structure analysis. Biochem Biophys Res Commun, 1990. **167**(2): p. 686-92.

10. Smith, R.D., et al., New developments in biochemical mass spectrometry: electrospray ionization. Anal Chem, 1990. **62**(9): p. 882-99.

11. Schar, M., K.O. Bornsen, and E. Gassmann, Fast protein sequence determination with matrix-assisted laser desorption and ionization mass spectrometry. Rapid Commun Mass Spectrom, 1991. **5**(7): p. 319-26.

12. Spengler, B., et al., Peptide sequencing by matrix-assisted laser-desorption mass spectrometry. Rapid Commun Mass Spectrom, 1992. **6**(2): p. 105-8.

13. Steiner, V., et al., Analysis of synthetic peptides using matrix-assisted laser desorption ionization mass spectrometry. Pept Res, 1992. **5**(1): p. 25-9.

14. Kuster, B. and M. Mann, Identifying proteins and post-translational modifications by mass spectrometry. Curr Opin Struct Biol, 1998. **8**(3): p. 393-400.

15. McLafferty, F.W., et al., Techview: biochemistry. Biomolecule mass spectrometry. Science, 1999. **284**(5418): p. 1289-90.

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

17. Fenn, J.B., et al., Electrospray ionization for mass spectrometry of large biomolecules. Science, 1989. **246**(4926): p. 64-71.

18. Karas, M. and F. Hillenkamp, Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem, 1988. **60**(20): p. 2299-301.

19. Tanaka, K., et al., Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of-Flight Mass Spectrometry. Rapid Commun. Mass Spectrom., 1988. **2**(8): p. 151-153.

20. Sze, S.K., et al., Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. Proc Natl Acad Sci U S A, 2002. **99**(4): p. 1774-9.

21. Syka, J.E., et al., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9528-33.

22. Keil-Dlouha, V.V., et al., Proteolytic activity of pseudotrypsin. FEBS Lett, 1971. **16**(4): p. 291-295.

23. Jekel, P.A., W.J. Weijer, and J.J. Beintema, Use of endoproteinase Lys-C from Lysobacter enzymogenes in protein sequence analysis. Anal Biochem, 1983. **134**(2): p. 347-54.

24. Schroeder, W.A., J.B. Shelton, and J.R. Shelton, An examination of conditions for the cleavage of polypeptide chains with cyanogen bromide: application to catalase. Arch Biochem Biophys, 1969. **130**(1): p. 551-6.

25. Wells, J.M. and S.A. McLuckey, Collision-induced dissociation (CID) of peptides and proteins. Methods Enzymol, 2005. **402**: p. 148-85.

26. Antoine, R., et al., Comparison of the fragmentation pattern induced by collisions, laser excitation and electron capture. Influence of the initial excitation. Rapid Commun Mass Spectrom, 2006. **20**(11): p. 1648-52.

27. Steen, H. and M. Mann, The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.

28. Taylor, J.A. and R.S. Johnson, Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 1997. **11**(9): p. 1067-75.

29. Ma, B., et al., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2337-42.

30. Frank, A. and P. Pevzner, PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem, 2005. **77**(4): p. 964-73.

31. Eng, J.K., A.L. McCormack, and J.R. Yates, An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom, 1994(5): p. 976-989.

32. Clauser, K.R., P. Baker, and A.L. Burlingame, Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem, 1999. **71**(14): p. 2871-82.

33. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999. **20**(18): p. 3551-67.

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

35. Colinge, J., et al., OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics, 2003. **3**(8): p. 1454-63.

36. Craig, R. and R.C. Beavis, TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004. **20**(9): p. 1466-7.

37. Geer, L.Y., et al., Open mass spectrometry search algorithm. J Proteome Res, 2004. **3**(5): p. 958-64.

38. Matthiesen, R., et al., VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J Proteome Res, 2005. **4**(6): p. 2338-47.

39. Tabb, D.L., C.G. Fernando, and M.C. Chambers, MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res, 2007. **6**(2): p. 654-61.

40. Tanner, S., et al., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 2005. **77**(14): p. 4626-39.

41. Tabb, D.L., A. Saraf, and J.R. Yates, 3rd, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem, 2003. **75**(23): p. 6415-21.

42. Hernandez, P., et al., Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics, 2003. **3**(6): p. 870-8.

43. Craig, R. and R.C. Beavis, A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2310-6.

44. Johnson, L.N. and D. Barford, The effects of phosphorylation on the structure and function of proteins. Annu Rev Biophys Biomol Struct, 1993. **22**: p. 199-232.

45. Hunter, T. and B.M. Sefton, Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. Proc Natl Acad Sci U S A, 1980. **77**(3): p. 1311-5.

46. Andersson, L. and J. Porath, Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. Anal Biochem, 1986. **154**(1): p. 250-4.

47. Posewitz, M.C. and P. Tempst, Immobilized gallium(III) affinity chromatography of phosphopeptides. Anal Chem, 1999. **71**(14): p. 2883-92.

48. Larsen, M.R., et al., Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol Cell Proteomics, 2005. **4**(7): p. 873-86.

49. Ficarro, S.B., et al., Niobium(V) oxide (Nb2O5): application to phosphoproteomics. Anal Chem, 2008. **80**(12): p. 4606-13.

50. Ficarro, S.B., et al., Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol, 2002. **20**(3): p. 301-5.

51. Salomon, A.R., et al., Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. Proc Natl Acad Sci U S A, 2003. **100**(2): p. 443-8.

52. Zhang, Y., et al., Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. Mol Cell Proteomics, 2005. **4**(9): p. 1240-50.

53. Beausoleil, S.A., et al., Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc Natl Acad Sci U S A, 2004. **101**(33): p. 12130-5.

54. McNulty, D.E. and R.S. Annan, Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. Mol Cell Proteomics, 2008. **7**(5): p. 971-80.

55. Schroeder, M.J., et al., A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. Anal Chem, 2004. **76**(13): p. 3590-8.

56. Zubarev, R.A., et al., Electron capture dissociation for structural characterization of multiply charged protein cations. Anal Chem, 2000. **72**(3): p. 563-73.

57. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res, 2003. **31**(13): p. 3635-41.

58. Beausoleil, S.A., et al., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol, 2006. **24**(10): p. 1285-92.

59. Olsen, J.V., et al., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell, 2006. **127**(3): p. 635-48.

60. Hornbeck, P.V., et al., PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics, 2004. **4**(6): p. 1551-61.

61. Peri, S., et al., Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res, 2003. **13**(10): p. 2363-71.

62. Barnidge, D.R., et al., Evaluation of a cleavable stable isotope labeled synthetic peptide for absolute protein quantification using LC-MS/MS. J Proteome Res, 2004. **3**(3): p. 658-61.

63. Gygi, S.P., et al., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol, 1999. **17**(10): p. 994-9.

64. Chakraborty, A. and F.E. Regnier, Global internal standard technology for comparative proteomics. J Chromatogr A, 2002. **949**(1-2): p. 173-84.

65. Kellermann, J., ICPL--isotope-coded protein label. Methods Mol Biol, 2008. **424**: p. 113-23.

66. Ross, P.L., et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.

67. Zhang, R., et al., Controlling deuterium isotope effects in comparative proteomics. Anal Chem, 2002. **74**(15): p. 3662-9.

68. Elias, J.E., et al., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods, 2005. **2**(9): p. 667-75.

Chapter 2

HTAPP: HIGH-THROUGHPUT AUTONOMOUS PROTEOMIC PIPELINE

2.1 INTRODUCTION

Dramatic progress has recently been made in expanding the sensitivities, resolving power, mass accuracy, and scan rates of mass spectrometers that can fragment and identify peptides through tandem mass spectrometry (MS/MS) [1-4]. Unfortunately, this enhanced ability to acquire proteomic data has not been accompanied by increased availability of tools able to assimilate, explore, and analyze these data efficiently. The typical proteomics experiment can generate tens of thousands of spectra per hour, and the use of multidimensional LC/MS, as with the MudPIT technique [5], can generate even larger datasets.

Computational tools for the collection and analysis of proteomic data lag far behind analytical methods for proteomic data creation [6]. In a typical experiment, collection and analysis of data is a fully manual process requiring repetitive and laborious sample- and data-processing steps with much unnecessary user intervention [6]. Proteomic datasets are expansive; adequate systems for the initial storage of proteomic data and its relationships to data from other external protein knowledge sources are inflexible and not integrated with the software used in data acquisition.

There are two options for handling the massive and diverse workflows in the modern proteomics lab: either provide a completely integrated software platform that is malleable to the users' needs, or provide independent software tools that require extensive user intervention to complete a total analysis of the data. Great progress has been made in providing independent software tools such as software focused on a single aspect of the proteomic pipeline. However, proteomic end users are left to fend for themselves in passing data amongst the various software tools and in modifying the individual software

tools to provide the processing and analysis needed for interpretation of their specific data. For example, one software tool is used for data acquisition, (such as Xcalibur or Analyst). A second tool interprets tandem mass spectra (such as X!Tandem [7, 8], Mascot [9], SEQUEST [10], OMSSA [11]) or statistical validation of database search results (such as Peptide/Protein Prophet [12], or Ascore [13]). A third tool provides quantitation of proteomic data (such as Xcalibur XDK, or MSQuant [14]), and a fourth provides a relational database for data warehousing (such as PRIME [15] or PeptideAtlas [16]) or a database graphical user interface for visual analysis of proteomic database search results (CPAS [17]). An assortment of web-based protein knowledge resources such as Swiss-Prot [18], HPRD [19], Genbank [20], OMIM [21], BLAST [22], IPI [23], and STRING [24] provide rich annotation of the proteins revealed in high-throughput proteomic experiments. These web-based metadata tools do not permit users to organize these external information sources relationally within the expansive proteomic datasets or to archive user observations. Although each of these tools provides essential functionality, they have not necessarily been engineered to adapt to diverse proteomic workflows or to work together efficiently.

Recent progress has been made in developing integrated systems for post-acquisition processing of data from high-throughput proteomic analysis. Notably, the Trans-Proteomic Pipeline [25] (TPP) integrates many critical aspects of post-acquisition proteomic analysis, including user initiated MS/MS sequence assignment, validation, quantitation and interpretation. To further expand the concepts driving the creation of workflow automation systems for proteomics such as TPP, we have now integrated

sample management, data acquisition, post acquisition analysis, and data visualization as integral components of a fully autonomous analysis pipeline called HTAPP.

2.2 MATERIALS AND METHODS

2.2.1 Overall Scheme

The overall scheme of HTAPP is illustrated in Figure 2.1. While each individual component of the integrated system can provide critical functionality independently, it is the interoperability of the components that provides a complete technology platform integrating data collection, storage, and visualization. In parallel with the development of HTAPP we have also developed a new relational database for proteomic analysis called PeptideDepot (see Chapter 3). HTAPP automatically directs the incoming data stream into PeptideDepot where a user may then interact with the processed proteomic data.

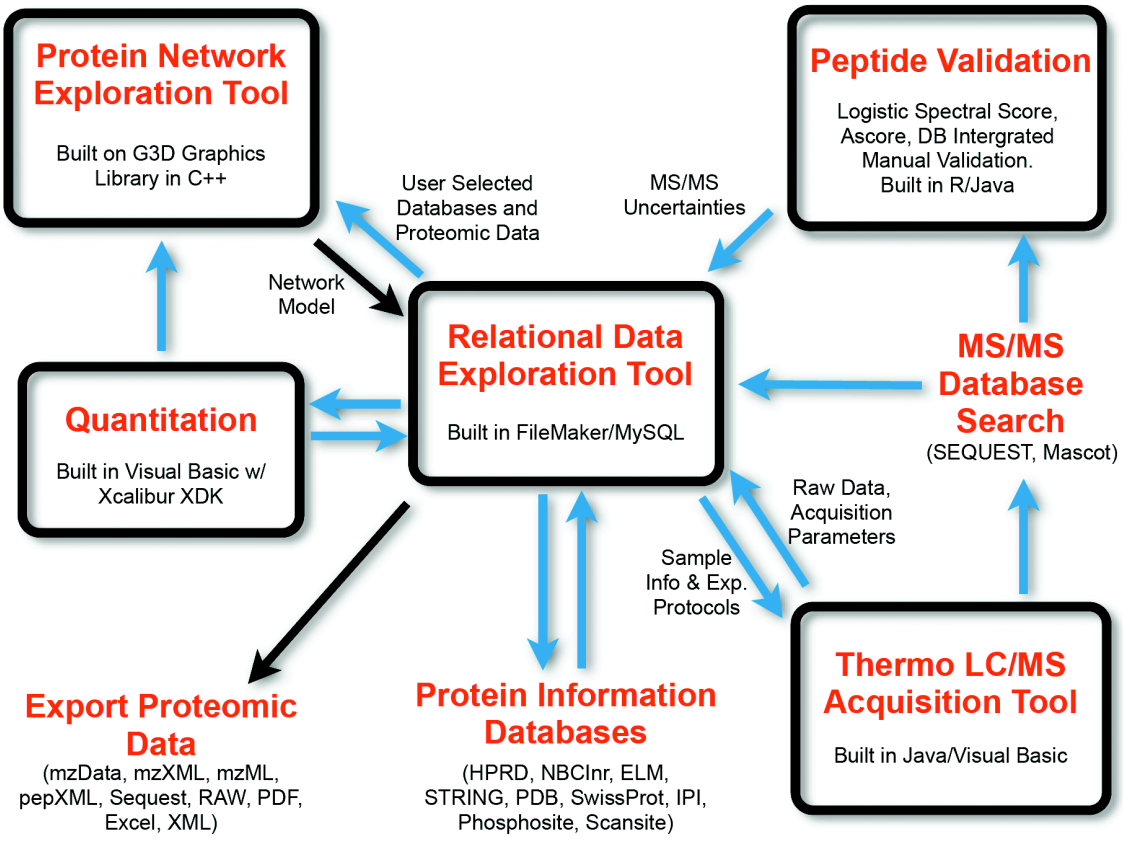


Figure 2.1: Overview of the HTAPP proteomic pipeline. HTAPP streamlines the collection and autonomous analysis of proteomic data. This system performs automated LC/MS data generation, identification, validation, quantitation, and integration with external protein information databases and enables protein network exploration. Blue arrows indicate actions that are performed automatically and black arrows describe tasks that require users' intervention.

2.2.2 Parallel Processing

To accelerate data processing and enhance system performance through parallel processing, the system components of HTAPP reside separately on several computers running Windows Server 2003 or Windows XP (Figure 2.2). The separate computers exchange data via TCP/IP protocol and Windows file-sharing network.

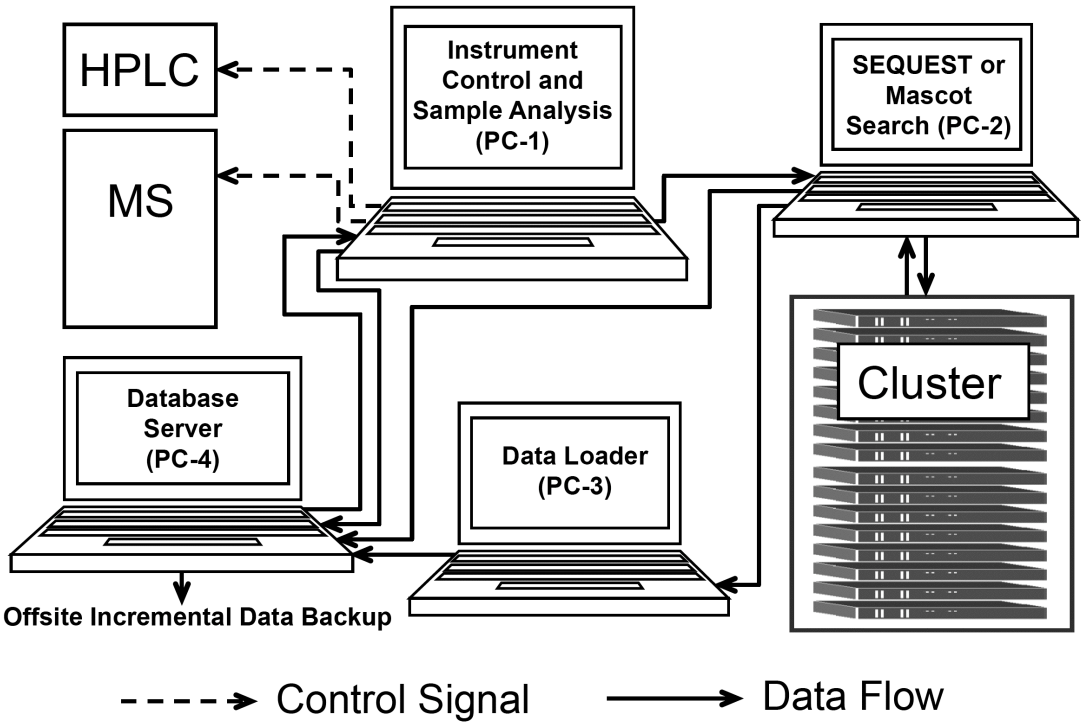


Figure 2.2: Diagram showing intercomponent data flow and communication scheme for HTAPP. HPLC: high performance liquid chromatography system; MS: mass spectrometer; PC-1: Data acquisition component; PC-2: transfer raw data files and perform SEQUEST or Mascot database search; Cluster: perform clustered SEQUEST or Mascot search; PC-3: autonomous post-acquisition analysis such as spectral validation, peptide quantitation, and upload proteomics data into PeptideDepot database; PC-4: database server (PeptideDepot) for visualization of proteomic data. Once data is deposited into PeptideDepot, it is incrementally backed-up offsite daily.

The modular design of HTAPP allows increased throughput as each component of the analysis workflow is performed simultaneously on separate computers. Through use of a distributed system, parallel processing enables the complete analysis of a proteomic data set within the acquisition time of the next proteomic sample. For example, an experiment containing 10,000 total MS/MS spectra in which ~1,000 spectra are high-quality (as defined by user determined thresholds) requires 1.5 hours to acquire the raw data on the mass spectrometer coupled to PC-1, 1 hour to perform a clustered SEQUEST search on PC-2 and the database search cluster, and 1.5 hours to complete the post-processing tasks including loading of data into the PeptideDepot relational database. Since SEQUEST

search and post-processing can be quite CPU-intensive, sequential processing of the data on a single computer requires approximately 4 hours per sample. However, with the distributed system the overall time is reduced to a total of 1.5 hours per sample.

2.2.3 Data Acquisition Software Module

An automated data acquisition tool developed in Microsoft Visual Basic 6.0 (VB6) runs on PC-1 to organize the predefined sample queue for analysis and to control a set of instrument manufacture software using Visual Basic (Figure 2.2, and Figure 2.3D). The extensibility of this tool is derived from flexible instrument control using Visual Basic SendKeys commands allowing the autonomous operation of any instrument control software. This central component of the automated acquisition of LC/MS data controls the unmonitored separation of peptides in, at most, three dimensions of chromatography and a simplified version has been described previously [26].

Here, we expand this data acquisition tool to integrate it within a data analysis pipeline that includes a relational database organized sample queue, MS/MS database searching, validation, and quantitation pipeline that automatically deposits the proteomic data and associated analysis within a relational database called PeptideDepot. An ODBC connection (Figure 2.3E) between the sample queue in PeptideDepot (Figure 2.3C) and the Visual Basic data acquisition software (Figure 2.3D) allows retrieval of selected sample information from the PeptideDepot database (FileMaker, version 9.0.3, FileMaker Inc., Santa Clara, CA). During the run, real-time instrument status information such as HPLC pressure profiles (Figure 2.4D), automated evaluation of selected ion chromatogram peak areas of peptides from standard mixes (Figure 2.4F), and screen

captures (Figure 2.4E) are archived in the MySQL (version 5.1.16-beta-nt; MySQL Inc., Cupertino, CA) component of PeptideDepot using a VB6 program. This data is available remotely through a website (Figure 2.4D-F) driven by Apache 2.2.4 (The Apache Software Foundation, Los Angeles, CA) and PHP 5.2.1 (http://www.php.net/).

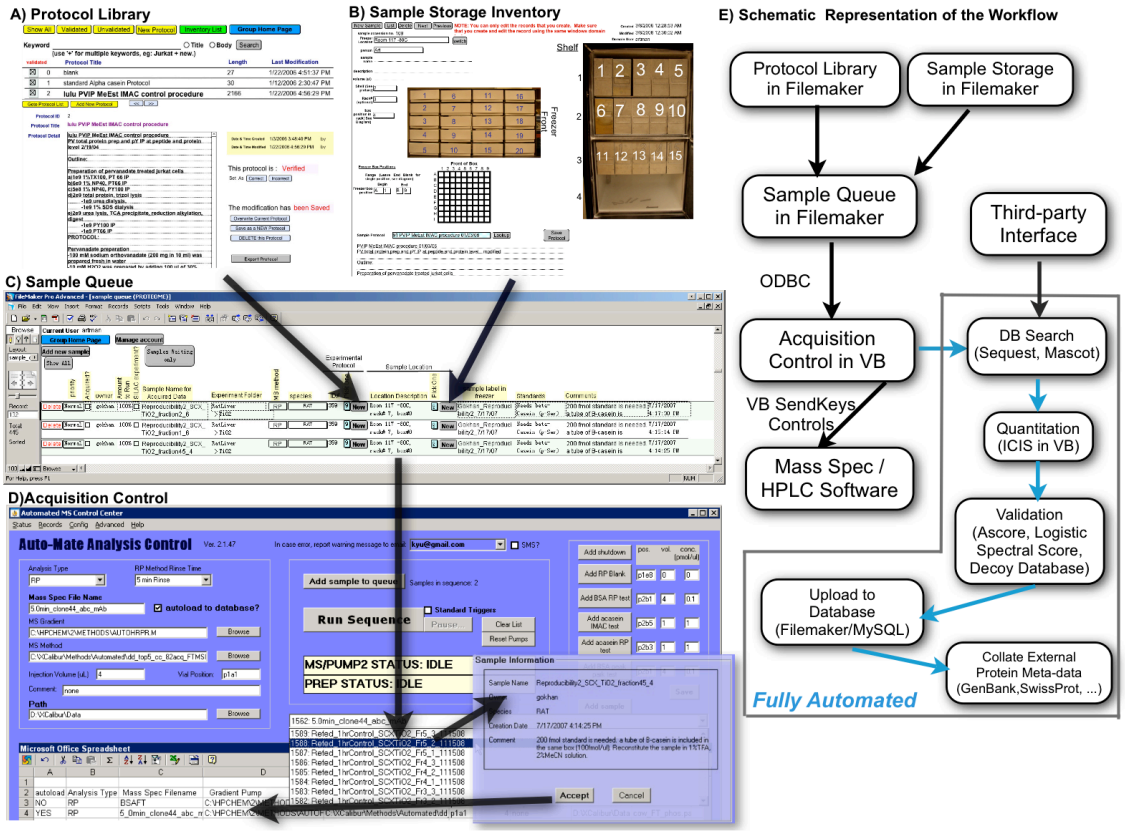


Figure 2.3: Integration among relational database component, data acquisition control, and fully autonomous post-acquisition analysis of HTAPP. A) A sample-generating user enters the protocol used in sample creation, B) the location of the sample, and any critical post-acquisition parameters into a C) sample queue located within our FileMaker database. D) The mass spectrometer operator then selects the sample for analysis with the acquisition control software component of our integrated system, which resides on the mass spectrometer control computer. All preferences for post-acquisition analysis, such as database search parameters and quantitation choices, are passed automatically to the acquisition control software from the sample queue and may be optionally modified by the instrument operator. When 'run sequence' is clicked, the acquisition control software communicates directly with data acquisition software provided by the instrument manufacturers via flexible VB SendKeys controls. E) Immediately after data acquisition is complete, the acquisition control software initiates automated data analysis, including MS/MS database searching, quantitation of relative peptide abundance, validation of peptide sequence assignments, loading of resulting data into FileMaker/MySQL, and caching of relationships between newly collected proteomic data and existing protein knowledge imported from external protein information databases and located internally within FileMaker/MySQL.

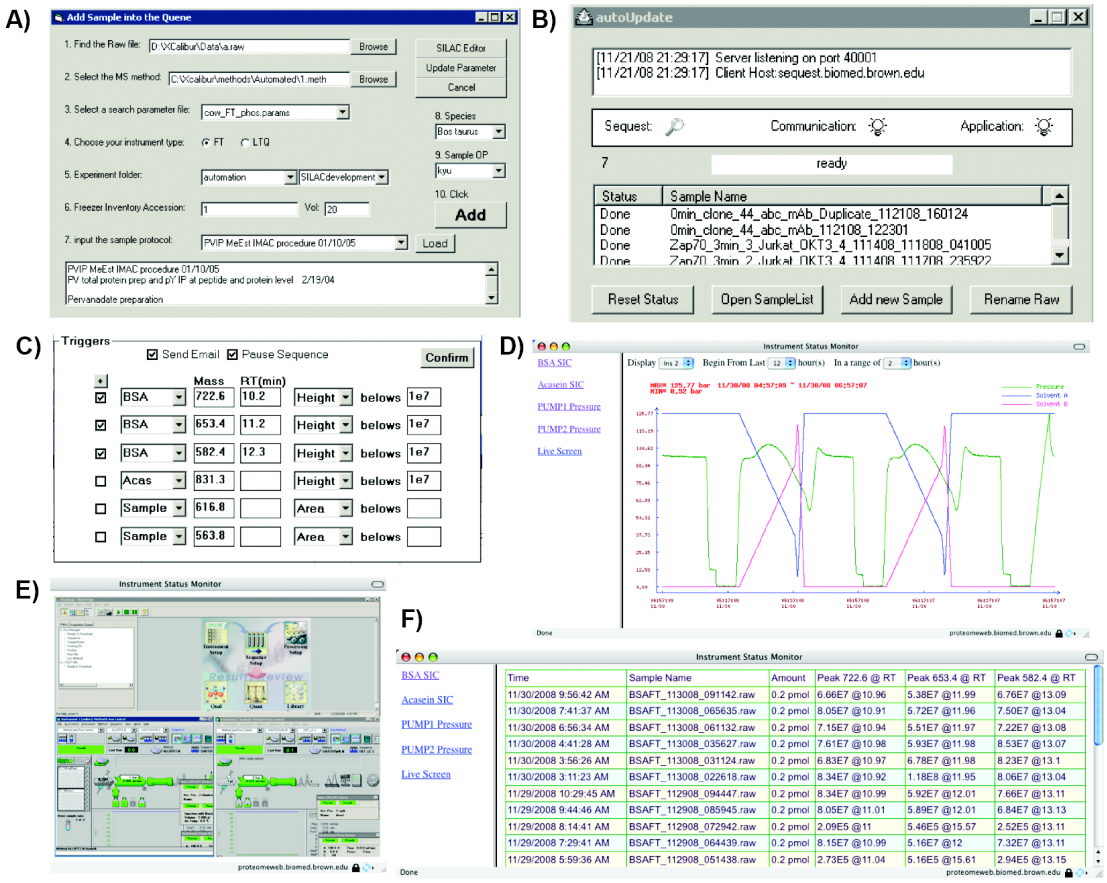


Figure 2.4: Automated analysis and troubleshooting within the HTAPP LC/MS data acquisition module. A) Designation of automated SEQUEST search and database deposition parameters for proteomic samples; B) Post-acquisition data pusher for initiation of autonomous post-acquisition analysis; C) Thresholds set for automated real-time monitoring of selected ion chromatogram peak heights or areas of bovine serum albumin (BSA), alpha casein peptides and user-selected masses in users' proteomic samples that triggers email alerts and/or halts the automated acquisition queue; D) Historical archive of HPLC gradient and pressure profile monitoring for multiple pumps displayed in a web browser; E) A webpage to monitor the live running status of LC/MS; F) Historical archive of three selected BSA peptide ion chromatogram peak areas from automated standard runs.

2.2.4 Automated Post-Processing Software Modules

Once a sample tagged as 'Autoload' is acquired on PC-1, a VB6 program running on PC-2 is notified via TCP/IP communication on port 40002 and transfers raw data files from PC-1 through Windows file sharing (Figure 2.2). MS/MS spectra are extracted from Thermo RAW files using extract_msn.exe (version 4.0; Thermo Scientific, Waltham, MA) or extracted from mzData, mzXML and mzML format using

ExtractMSMS.jar (in-house developed in Java 1.6.0; Sun Microsystems, Santa Clara, CA) to generate DTA files. A SEQUEST cluster (version 27; Thermo Scientific) or Mascot cluster (version 2.2.1; Matrix Science) MS/MS database search is initiated via a networked computer cluster.

After completion of SEQUEST or Mascot searching, proteomic data is transferred using Windows file sharing to a third computer, PC-3, which is reserved for a variety of post-processing tasks (Figure 2.2). On PC-3, a variety of independent calculations are performed on the proteomic data. A VB6 program called "AutoLoad" orchestrates the initiation and transfer of data amongst these separate software tools. A peptide quantitation tool and SILAC calculation tool are created in Visual Basic 6 using the Xcalibur XDK. A phosphosite localization tool that calculates Ascore as described previously [13] is written in Java 1.6.0, and MS/MS validation tool implementing a new user-trainable logistic regression algorithm that more than doubles peptide identifications at a user selected false discovery rate compared to XCorr [27] is implemented in R 2.4.1 (The R Foundation, http://www.r-project.org/). Once the calculations are finished, proteomic data are immediately uploaded to a FileMaker/MySQL relational database called PeptideDepot hosted on the remote server PC-4 using FileMaker script and PHP5 scripts. The proteomic data is then accessible from a graphical FileMaker client (version 9.0.3) running on Mac or Windows. The database files are synchronized daily without user intervention to an offsite server for the incremental backup using either the commercial software tool Retrospect 7.5 (EMC Insignia; Pleasanton, CA) or the Carbonite backup service (http://carbonite.com; Boston, MA).

2.3 RESULTS

2.3.1 Automated Acquisition Control

To create a robust infrastructure capable of high-throughput analysis of proteomic samples, we sought tight integration between the bioinformatic tools used in analyzing proteomic data and the software involved in acquiring mass spectral data. This system can flexibly automate projects ranging from simple LC/MS of in-gel digested proteins to more complex proteomic analyses, such as 2D nano-LC/MS experiments or protein post-translational modification analyses.

2.3.2 Sample Queue Management and Automated Workflows

A sample queue capability within the FileMaker component of PeptideDepot relational database integrates sample creation, and metadata annotation with data acquisition control and automated post-acquisition analysis (Figure 2.3A-D; Figure 2.4A,B). This system provides unparalleled flexibility to the user by 1) letting any user tailor the sample queue in FileMaker for automation of any lab-specific post-acquisition analysis task or association of any experimental meta-data with the nascent proteomic data, 2) providing flexible control of any mass spectrometer using a system that employs Visual Basic SendKeys to manage data acquisition software from any instrument manufacturer, and 3) providing an array of choices for post-acquisition analysis for the automated or manual interpretation of proteomic data.

The laboratory information management system (LIMS) components of the PeptideDepot database are created in the user-friendly FileMaker environment, allowing proteomic end-users to tailor the associated fields and layouts to their specific needs

(Figure 2.3A-C). For example, users wanting to store a new piece of information within the system to be automatically associated with the analyzed proteomic data may quickly add a field for this data in FileMaker and position it precisely within user-defined layouts with FileMaker's WYSIWYG layout tools (such as illustrated for the protocol library and sample storage inventory in Figure 2.3A-B). With this flexibility, the end-user need not wait for a programmer or database engineer to add the desired functionality; it may be implemented directly.

2.3.3 Automation of Post-Acquisition Data Analysis

Although sample metadata may vary dramatically from lab to lab, the processing of proteomic data after acquisition most commonly involves some combination of database searching, quantitation, validation of database search results, and storage of proteomic data within a relational database. A variety of software tools are used in each step of this standard analysis pipeline (summarized in Figure 2.1). For database searching, our automated system currently supports SEQUEST, Mascot, or any other algorithm that exports to pepXML. For quantitation, our automated system currently uses the ICIS algorithm available in the Xcalibur XDK to calculate peak areas for label free or isotopic labeling methods such as SILAC from any Thermo Scientific Xcalibur (RAW) file. For validation, our system currently automates the analysis of reversed database searches [28], performs peptide validation using a recently developed logistic spectral score that more than doubles peptide yield at a fixed FDR [27], and phosphorylation site localization using the Ascore algorithm [13]. Our relational database PeptideDepot also provides unique tools, namely SpecNote for database-integrated manual validation and

annotation of spectra [27]. Our current system provides for unmonitored import of proteomic data and proteomic analyses into our flexible PeptideDepot relational database that utilizes a FileMaker generated user interface.

2.3.4 Flexible Workflow Support

Although the software tool that performs automated data acquisition currently incorporates a Thermo Scientific hybrid Linear Ion Trap–Fourier Transform mass spectrometer (LTQ-FTICR) and Agilent 1100 HPLC pumps, our control software is adaptable to any mass spectrometer and chromatography system through the use of flexible Visual Basic SendKeys controls [29]. In its current implementation, SendKeys works through Xcalibur and Chemstation software to control the automated acquisition of data. Using SendKeys controls, our software sends keyboard commands to any currently running software. By using SendKeys, control of additional mass spectrometer data acquisition software systems can be rapidly implemented to provide critically important extensibility to our automated platform.

HTAPP also supports the analysis of any additional mass spectrometer MS/MS data that may be converted to the standard proteomic data formats, *i.e.* mzData [30], mzXML [31] and mzML [32] (Figure 2.5). Tools to convert manufacturer-specific raw data to standard formats are publicly available (http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML). For Thermo Scientific RAW files, the analysis pipeline is fully automated. If a user desires to analyze data from other types of mass spectrometers, the user first converts the data to either mzData, mzXML or mzML format using publicly available software prior to autonomous

analysis through HTAPP. We have implemented a Java program in HTAPP to convert MS/MS spectra from standard formats and initiate autonomous data analysis. This software was verified with publicly available proteomic datasets acquired on Agilent, LCQ-Deca, LTQ and QSTAR mass spectrometers [33].

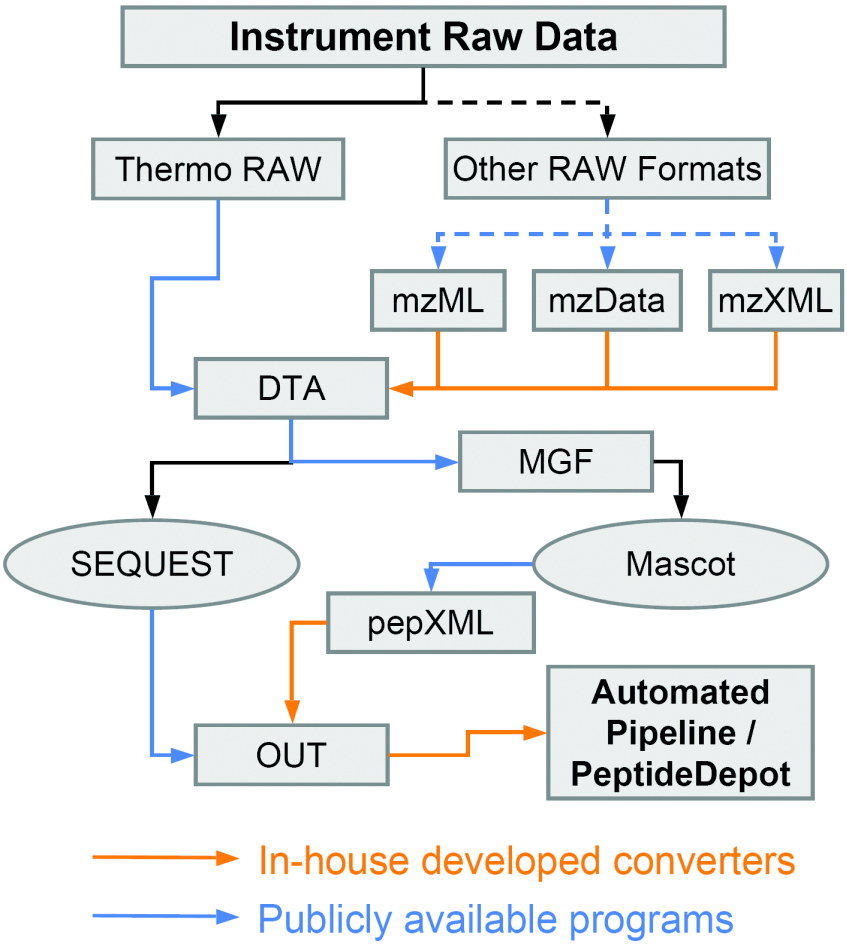


Figure 2.5: Flexible workflows through support of standard proteomic data exchange formats. In the figure, mzML, mzXML and mzData are standard XML formats for MS/MS data. DTA is the generic format for SEQUEST input. MGF is the Mascot Generic Format. pepXML is the standard XML format for database search output. OUT is the generic format of SEQUEST search output. Conversions indicated with solid arrows are accomplished autonomously within HTAPP while the dashed lines indicate the tasks that require user intervention. Post-processing of additional mass spectrometer specific raw data formats is provided through support of mzML, mzXML and mzData formats. Database search engines beyond SEQUEST and Mascot can be implemented by converting DTA to a particular format for that search engine and exporting search output in pepXML format, as is illustrated for Mascot here.

After data acquisition, the peptide sequences are assigned through a SEQUEST or Mascot cluster, peptides quantitated, uncertainties of peptide and phosphorylation site placement are accessed, and proteomic data are deposited into a networked relational database (Figure 2.3E). If a user's workflow includes additional analysis tasks beyond the core functionality already available within HTAPP, these additional calculations may be automated through FileMaker scripts which export the proteomic data in standard formats, trigger external analysis software, and import the analysis results back into the PeptideDepot database into user-defined fields that are displayed on user-configured layouts.

2.3.5 Sample Tracking Database and Protocol Library

We have also created a sample tracking database and protocol library (Figure 2.3A,B) that organize information about sample preparation and storage and associate this information with the nascent proteomic data. This tool enhances the ability to find correlations between proteomic results and the conditions used to prepare and store samples while facilitating post-acquisition analysis by specification of data processing parameters prior to data acquisition. These tools are dynamically integrated within our data acquisition and automation tools to facilitate the automation and documentation of samples awaiting proteomic analysis. By requiring the entry of sample protocols before data acquisition, critical experimental conditions and metadata are captured, organized, and associated with complex proteomic datasets. Also, the protocol library allows assimilation of all protocols used in the lab within a lab-based relational database and

provides a mechanism by which protocols can be reviewed and optionally approved by other researchers.

2.3.6 Automated Monitoring and System Troubleshooting

To promote efficient troubleshooting of fluctuations in system performance, the automated data acquisition includes the capability to store and analyze metadata captured during spectral acquisition in a fully automated fashion. Information such as the pressure profiles and chromatography gradients are all automatically archived in the MySQL component of the PeptideDepot relational database that is linked to the raw data and SEQUEST results and accessible through a web-based PHP interface (Figure 2.4D). Selected ion chromatogram (SIC) peak areas of either Bovine Serum Albumin (BSA) or α-casein peptides from automated standard runs, or of user-selected standard peptides incorporated into user samples, are monitored automatically. If any selected peptide falls below a user-defined threshold, the operator is optionally alerted via email or instant SMS and the acquisition queue can be set by the user to halt (Figure 2.4C). A user may also explore all the historical BSA and α-casein SIC data acquired on the instrument in an interactive web browser layout driven by PHP (Figure 2.4F) or in a VB6 program to track and troubleshoot instrument sensitivity over time. Remote access capabilities allow any operator to monitor the status of the system in real time (Figure 2.4E) and to control the system through an encrypted network connection.

2.3.7 Relational Database for Proteomic Data Exploration

Proteomic results are automatically imported to a networked relational database called PeptideDepot (see Chapter 3). Tight integration of external protein information sources is a critical aspect of this system. Once newly acquired data are deposited into the PeptideDepot database, many data-mining calculations are triggered automatically by querying externally available protein information databases such as PDB [34], IPI [23], HPRD [19], Swiss-Prot [35], STRING [24], Phosphosite [36] and Scansite [37] by peptide sequence across locally cached databases. All possible protein names associated with a given peptide sequence are collated from the locally cached external protein information databases. This capability overcomes the limitation of alternative protein naming by allowing for users to "deep search" the data sets across an index of all possible protein names in every database.

After automated analysis and deposition of the data within PeptideDepot, users may explore the data with flexible FileMaker WSIWYG layouts. PeptideDepot features an extensive collection of predefined data filters that enable users to limit false-discovery rates estimated by reversed database search while focusing on specific peptide qualities such as tyrosine phosphorylation, etc. Comparative analysis views, useful in comparing peptides observed in different cellular states such as disease versus healthy tissue, are provided to facilitate quantitative comparison among samples using either label-free or stable-isotope incorporation quantitation strategies such as SILAC.

2.4 DISCUSSION

One of the largest impediments to truly high-throughput proteomic methods is the lack of automation after the acquisition of spectra and lack of capture of critical acquisition-specific metadata. In addition, there is a fundamental need not only to acquire data more quickly but also to increase the quality of data acquired. An ideal high-throughput proteomic pipeline would provide for the thorough documentation of a sample's provenance: the protocol used in sample preparation, sample storage information, environmental conditions such as temperature and humidity during the analysis, and HPLC gradients and pressure profiles.

2.4.1 Integrative Approach to Proteomic Analysis

One of the fundamental goals of the work described here is to provide truly high-throughput multidimensional acquisition of spectra coupled to automated database searching, data archiving, data filtering, visualization, analysis, quantification, and statistical validation of spectra. The software described here uses an integrative approach in which all information concerning a proteomic experiment is archived automatically along with the raw data and database assignments. All components of analysis are integrated within a lab-centric relational database. Capturing a myriad of experimental metadata in addition to spectral acquisition enables the organization and documentation of complex experiments and facilitates troubleshooting. Unlike other currently available proteomic software, our integrated platform utilizes a sample queue in which post-processing parameters and user-provided proteomic sample annotation are passed directly to data acquisition control software and are associated automatically with proteomic data as it is collected and processed within a lab's relational database. This tight integration

greatly increases efficiency by automating labor-intensive post-processing tasks and reduces the chances that critical connections between newly collected proteomic data and experimental metadata will be lost.

This work provides an integrated yet extensible technology platform for the automated processing, storage, and visual analysis of expansive proteomic datasets. Instead of trying to patch together a variety of preexisting software tools that fit together awkwardly, match analytic needs only marginally, and lack critically important functionality, we have created from the ground up an optimized set of integrated tools that provides automated acquisition, processing and visual analysis of proteomic data. Although many aspects of our software implementations are both unique and essential for a thorough analysis of these types of data, the main novelty of our approach is the direct software integration of the collection, quantitative processing, and visual analysis of proteomic data. No publicly available software tool currently available provides this level of integration. Current proteomic end-users must either develop their own proteomic pipeline software systems in each lab or else perform tedious data manipulation steps manually to extract biological meaning from the immense datasets.

2.4.2 Extensibility

The HTAPP software is designed to provide critical flexibility and functional extensibility for users to implement alternative proteomic workflows as needed. For example, if a user wants to acquire data on a new instrument, the SENDKEYS commands in our software could be changed to the series of timed keystrokes necessary to control

any instrument manufacturers data acquisition software, without altering the whole proteomic workflow.

Although the analysis of MS/MS data acquired on ThermoScientific mass spectrometers is fully automated, data generated by other instruments can be processed manually within HTAPP as well by converting the data to the standard mzML, mzXML or mzData formats before automated HTAPP analysis.

To support flexible expansion for future software to interact with the automated pipeline, samples awaiting analysis reside in two independent flat-file formatted sample queues. The first sample queue resides on the data acquisition component (PC-1; Figure 2.2) while the second queue resides downstream of the database search component on the data loader (PC-3; Figure 2.2). By adding, removing, or altering the text formatted sample queues, a user can integrate their own software within the HTAPP pipeline.

To incorporate a new database search engine such as X!Tandem for MS/MS interpretation, the proteomic researcher only need to configure the database search program to export the results in the standard pepXML [25] format and trigger existing pepXML import scripts that are already available in HTAPP (Figure 2.5). Once imported to FileMaker, the parsed database search results would be integrated into user-defined flexible layouts.

To accomplish any additional post-acquisition data analysis task, the sample queue table within FileMaker has a unique counter field that is transferred throughout the data analysis pipeline and stored with the analyzed proteomic data. Using this counter field, proteomic end users may add any post acquisition preferences to the sample queue, and optionally trigger the execution of external software tools using FileMaker scripts that

export the proteomic data from PeptideDepot, trigger the external program and import the results of the external analysis back into FileMaker for display on user-defined custom layouts. For fully automated post-acquisition analysis, the existing FileMaker data import script can then optionally trigger these external calculations.

2.4.3 Overall Benefit

The laboriousness of current proteomics software manual implementations distracts the proteomics investigator from the biological meaning of the data, leading to the all-too-frequent deposition of data into the scientific literature with minimal biological or clinical interpretation. Instead of treating individual steps in the proteomic pipeline as separate events whose integration depends on end-user intervention, we let the user focus on the interpretation of the data through automation of routine data manipulations and caching of comparisons between newly collected proteomic data and external bioinformatic resources within a lab-based relational database.

2.5 REFERENCE

1. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry.* J Mass Spectrom, 2001. **36**(8): p. 849-65.

2. Schwartz, J.C., M.W. Senko, and J.E. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* J Am Soc Mass Spectrom, 2002. **13**(6): p. 659-69.

3. Syka, J.E., et al., *Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications.* J Proteome Res, 2004. **3**(3): p. 621-6.

4. Yates, J.R., et al., *Performance of a linear ion trap-Orbitrap hybrid for peptide analysis.* Anal Chem, 2006. **78**(2): p. 493-500.

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

6. Topaloglou, T., *Informatics solutions for high-throughput proteomics.* Drug Discov Today, 2006. **11**(11-12): p. 509-16.

7. Craig, R. and R.C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2310-6.

8. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.

9. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

10. Eng, J.K., A.L. McCormack, and J.R. Yates, *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.* J Am Soc Mass Spectrom, 1994(5): p. 976-989.

11. Geer, L.Y., et al., *CDART: protein homology by domain architecture.* Genome Res, 2002. **12**(10): p. 1619-23.

12. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.

13. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nat Biotechnol, 2006. **24**(10): p. 1285-92.

14. Andersen, J.S., et al., *Proteomic characterization of the human centrosome by protein correlation profiling.* Nature, 2003. **426**(6966): p. 570-4.

15. Ulintz, P.J., et al., *4th Siena 2D Electrophoresis Meeting*. 2000, Siena, Italy.

16. Desiere, F., et al., *Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.* Genome Biol, 2005. **6**(1): p. R9.

17. Cottingham, K., *CPAS: a proteomics data management system for the masses.* J Proteome Res, 2006. **5**(1): p. 14.

18. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.

19. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

20. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.

21. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.

22. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

23. Kersey, P.J., et al., *The International Protein Index: an integrated database for proteomics experiments.* Proteomics, 2004. **4**(7): p. 1985-8.

24. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Res, 2005. **33**(Database issue): p. D433-7.

25. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Mol Syst Biol, 2005. **1**: p. 2005 0017.

26. Ficarro, S.B., et al., *Automated immobilized metal affinity chromatography/nano-liquid chromatography/electrospray ionization mass spectrometry platform for profiling protein phosphorylation sites.* Rapid Commun Mass Spectrom, 2005. **19**(1): p. 57-71.

27. Yu, K., et al., *Integrated platform for manual and high-throughput statistical validation of tandem mass spectra.* Proteomics, 2009. **in press**.

28. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

29. Cao, L., K. Yu, and A.R. Salomon, *Phosphoproteomic analysis of lymphocyte signaling.* Adv Exp Med Biol, 2006. **584**: p. 277-88.

30. Orchard, S., et al., *Current status of proteomic standards development.* Expert Rev Proteomics, 2004. **1**(2): p. 179-83.

31. Pedrioli, P.G., et al., *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol, 2004. **22**(11): p. 1459-66.

32. Deutsch, E., *mzML: a single, unifying data format for mass spectrometer output.* Proteomics, 2008. **8**(14): p. 2776-7.

33. Klimek, J., et al., *The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools.* J Proteome Res, 2008. **7**(1): p. 96-103.

34. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

35. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.

36. Hornbeck, P.V., et al., *PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation.* Proteomics, 2004. **4**(6): p. 1551-61.

37. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.* Nucleic Acids Res, 2003. **31**(13): p. 3635-41.

Chapter 3

PEPTIDEDEPOT: FLEXIBLE RELATIONAL DATABASE FOR VISUAL ANALYSIS OF QUANTITATIVE PROTEOMIC DATA AND INTEGRATION OF EXISTING PROTEIN INFORMATION

3.1 INTRODUCTION

Recent exciting progress in the development of new instrumentation and high throughput proteomic methods has led to a landslide of proteomic data that needs to be analyzed and explored efficiently [1-4]. A single LC/MS experiment potentially generates thousands of peptide-specific spectra, and the use of multidimensional separations, such as in the MudPIT technique [5], generates even larger datasets.

Quantitative proteomics experiments facilitate the analysis of protein levels between various samples, such as between diseased and normal tissue, or across time points in the analysis of cellular signaling associated with receptor stimulation or drug treatment. Many quantitative proteomic methods have been developed and reviewed recently [6]. Stable isotopes have been incorporated within peptides either through Stable Isotope Labeling by Amino acids in Cell culture (SILAC) [7] or through direct labeling of peptides with Isotope Coded Affinity Tags (ICAT) [8], isobaric Tags for Relative and Absolute Quantitation (iTRAQ) [9], or other chemical labeling techniques [10]. An alternative 'label-free' approach to relative quantitation employs normalization of peptide peak areas against a co-purified exogenous standard peptide [11]. Visual representation of quantitative proteomic data such as with a heatmap allows for rapid comparison between various cellular states but is not a feature available in existing software tools. Instead current tools such as Bioworks (Thermo Fisher, San Jose, CA) or Mascot (Matrix Science, Boston, MA) convey quantitative data as columns of numbers or ratios if at all. Rapid comparison between samples is not possible with this type of presentation. Collation of replicate quantitative datasets and efficient access to the underlying selected ion chromatograms through heatmap navigation is not currently supported in existing

software tools. Instead, a researcher hoping to validate the quantitative measurement must manually query each peptide in a separate piece of software such as Xcalibur or Analyst, decreasing efficiency and increasing the chance for error.

Proteomic journal-mandated warehousing of published proteomic data [12] has led to the creation of a variety of relational databases such as Peptide Atlas [13], Human Proteinpedia [14], CPAS [15], PRIME [16], and PRIDE [17]. These existing databases provide critically important means of distribution of data between proteomic investigators but are not necessarily designed for direct proteomic end-user modification for lab-specific proteomic workflows in labs without dedicated programmer support. The general proteomic workflow is greatly diversified among labs and flexibility in the presentation, statistical analysis, quantitation, and filtering of the proteomic data and integration of proteomic data with existing protein information sources is a critically important component of high throughput proteomic workflows prior to publication. Here we introduce a flexible proteomic data warehousing and analysis repository, named PeptideDepot, to provide users with the critical information and quantitative analysis of acquired data, as well as a flexible interface for implementation of alternative data representations. The integration of FileMaker's WYSIWYG layout and schema editors, transparent integration with data warehoused in a MySQL relational database, and cached external protein information databases allows the relational database to become integrated into the workflows of a broader range of proteomic labs.

3.2 MATERIALS AND METHODS

Data can be loaded into PeptideDepot database manually through a user interface integrated directly within the PeptideDepot (Figure 3.1, Figure 3.3A) or it can be integrated directly with other software tools such as the High Throughput Autonomous Proteomic Pipeline (HTAPP) used in automated LC/MS data acquisition and troubleshooting, and autonomous post-acquisition analysis of proteomic data. Data generated manually by a proteomic researcher or autonomously using HTAPP are loaded into the PeptideDepot database. Supported input formats for MS/MS data and database search results in Peptide Depot include Thermo Raw, SEQUEST DTA/OUT, standard mzData [18], mzXML [19], mzML [20] and pepXML [21], providing support for any instrument or any database search method supporting standard proteomic formats.

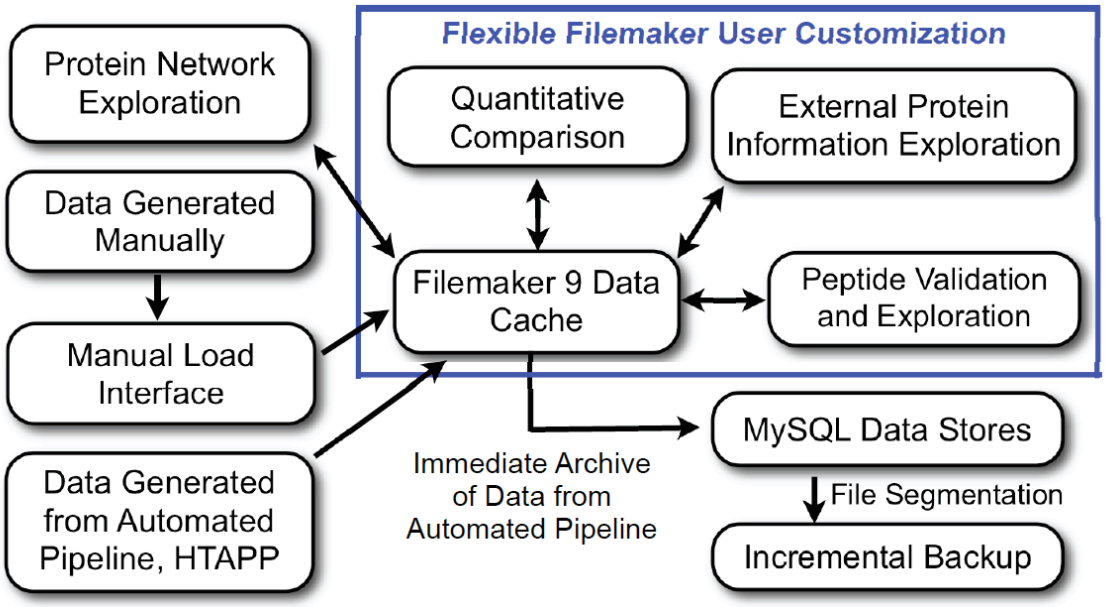


Figure 3.1: Design of data visualization tool providing flexible user customization of data representations, calculations, and secure proteomic data storage. Unlike other proteomic database software, proteomic end-users can customize the data representations and calculations within FileMaker without the need for dedicated programming staff.

For calculation of selected ion chromatogram peak areas for label free or stable isotope labeling methods, PeptideDepot utilizes either the Xcalibur XDK within a newly created Visual Basic program for the analysis of ThermoFisher .RAW files or the existing software tool ProteinQuant [22] for the extraction of quantitative data from any mass spectrometer data file that supports the standard proteomic data formats mzXML and mzData.

The PeptideDepot relational database consists of MySQL (version 5.1.16-beta-nt; MySQL Inc., Cupertino, CA) and FileMaker (version 9.0.3; FileMaker Inc., Santa Clara, CA) tables, which are transparently integrated for the proteomic end-user, using FileMaker's external SQL sources capabilities (Figure 3.2). Data may also be accessed from the FileMaker and MySQL tables through ODBC connection or accessed directly using PHP or Java. A set of graphical layouts within FileMaker allows the user to explore the proteomic data and associated experimental meta-data, as well as to export those data into Excel or PDF files with a format compliant with guidelines set by peer-reviewed proteomics journals. FileMaker scripts provide a simple and flexible way to manipulate the data stored within the database using intuitive FileMaker scripting language. Authentication of users is managed internally within the database or externally via a Microsoft Windows domain controller and is transparent to networked Windows clients logged into a Windows domain. The database may be accessed via a web page with any operating system or via a standalone FileMaker client in Windows or Mac OS X. In addition to the FileMaker client software, the data are accessible via ODBC, JDBC or the FileMaker PHP API. A user can either display a single proteomic experiment or make quantitative comparisons amongst multiple proteomic experiments.

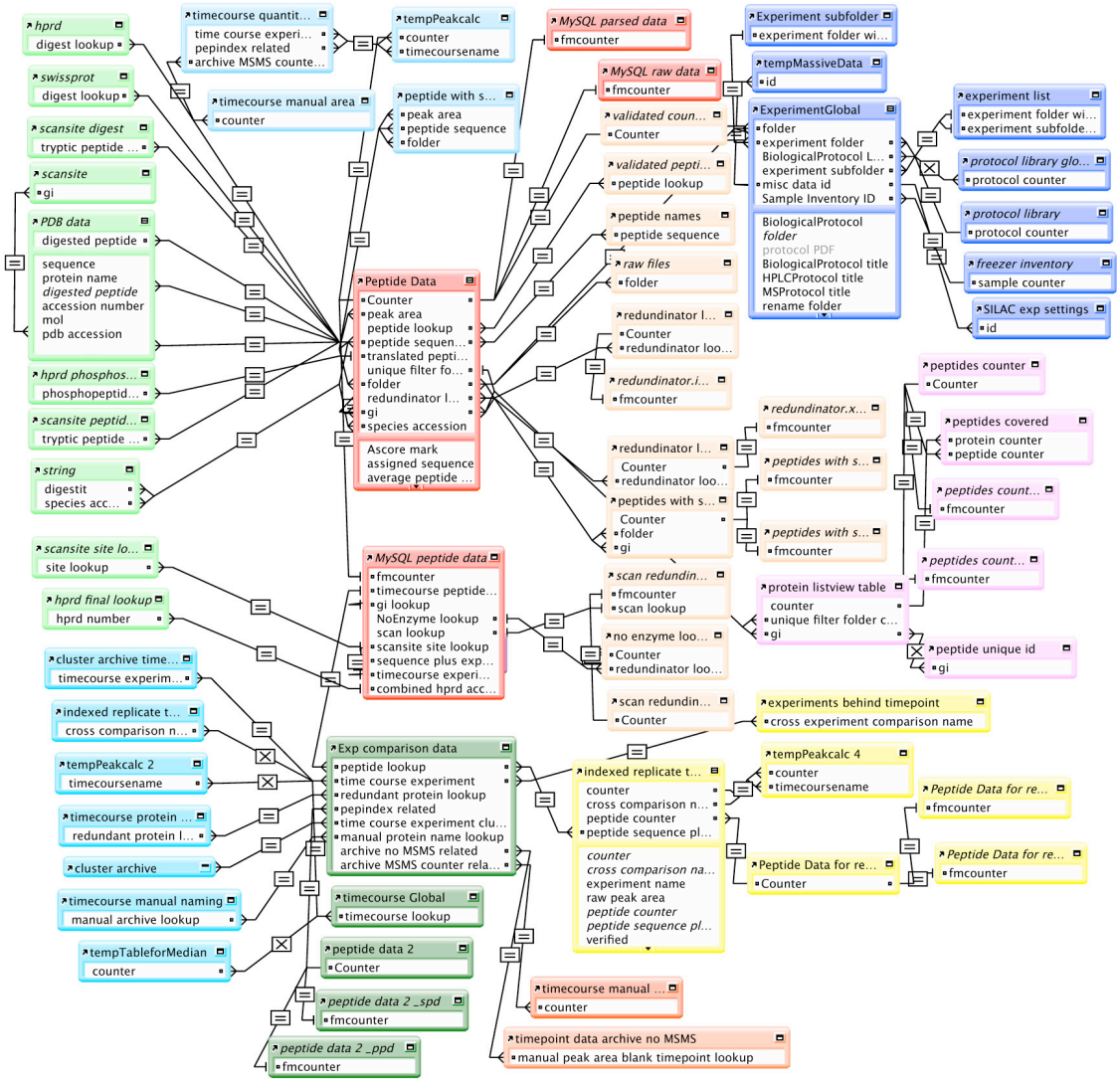


Figure 3.2: Entity-relationship diagram for PeptideDepot. Data warehoused in MySQL (labeled MySQL and colored red) is relayed transparently to FileMaker using its external SQL source functionality.

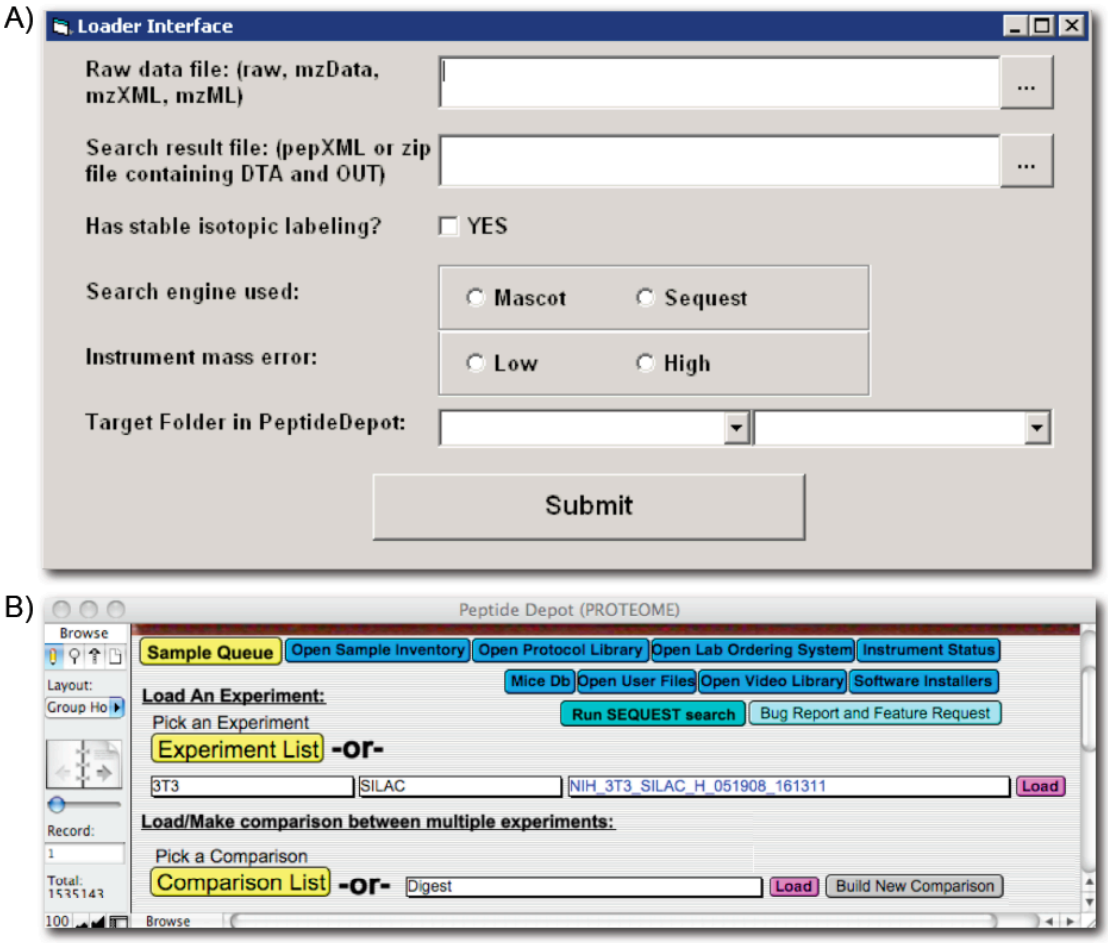


Figure 3.3: An intuitive user interface to A) manually load LC/MS experiment data into PeptideDepot, and B) explore previously loaded proteomic data and related resources.

3.3 RESULTS AND DISCUSSION

Our networked proteomic relational database PeptideDepot serves as both a data repository and a means of visualizing incoming data (Figure 3.1). A database homepage gives users an array of options for exploring proteomic data or customizable lab-related resources (Figure 3.3B). The user may explore the sample inventory database, and an experiment protocol library.

3.3.1 Quantitative Comparison Analysis Tools

PeptideDepot provides a unique intuitive way for quantitative comparison amongst previously loaded datasets using either label-free or stable-isotope labeling strategies as illustrated in Figure 3.4. These quantitative data representations allow for the comparison of peptide levels in different experimental conditions such as through a timecourse of cellular stimulation or to compare diseased to normal cells. Such a quantitative heatmap view of the data mimics similar representations that are typical in transcriptional profiling, using colors to indicate the relative abundance of a given peptide under various conditions.

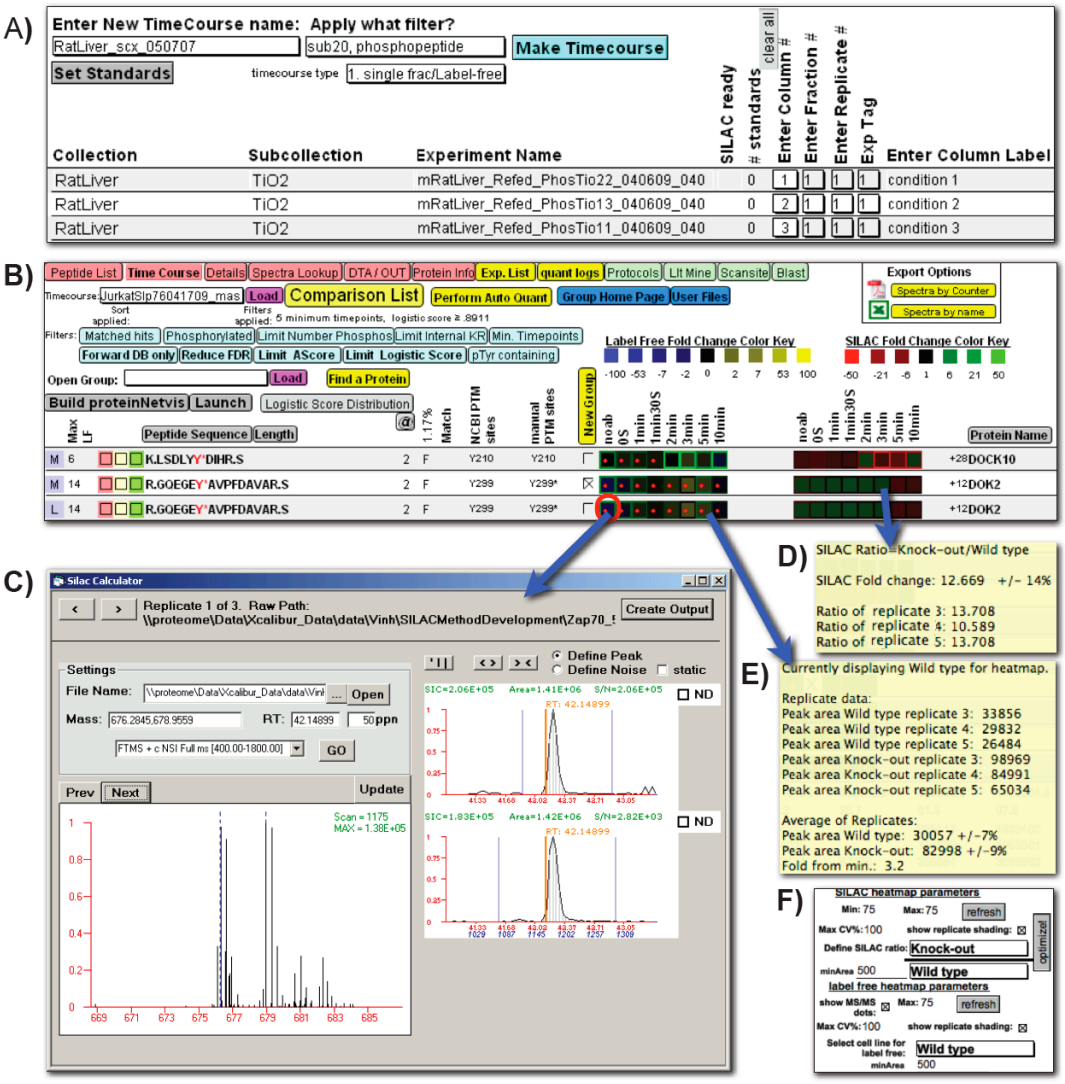


Figure 3.4: Quantitative comparison between multiple proteomic experiments within FileMaker, with heatmap navigation of underlying quantitative proteomic data. A) A FileMaker layout guides users through the creation of a new comparison amongst selected experiments; B) The heatmap is used for rapid quantitative comparison and as a navigational tool for validation of the quantitative data. Blue-Yellow label-free heatmap (Left) visualizes the quantative change across several cellular states and Red-Green ratio heatmap (Right) visualizes the change in ratio between two stable-isotope labeled samples. If replicate data is used to generate the heatmaps, the average values are presented while the error amongst the replicates is portrayed by a colored outline around each heatmap square. The intensity of this outline correlates to the magnitude of replicate error. Red dots indicate whether an MS/MS identification was captured for each peptide. A set of filters is available for the user to narrow down the whole dataset and focus on interesting proteins; C) Manual inspection and adjustment of the underlying peak area calculation accessed by clicking any heatmap square. Orange vertical lines represent the position of acquired MS/MS spectra for the selected peptide (including redundant MS/MS spectra for the same peptide), allowing selection of the correct peak. Both SIC and profile MS scan is shown for stable-isotope labeled data. The user may tab through the replicate spectra underlying the average value used to calculate the heatmap square color; D) and E) Hovering the mouse over a label-free heatmap or stable-isotope ratio heatmap square reveals the detailed underlying quantitative data including replicate peak areas and standard deviations; F) The user can adjust which data is portrayed in each of the heatmaps and adjust heatmap parameters.

A 'make comparison' layout is provided in FileMaker to build a quantitative comparison heatmap amongst selected experiments (Figure 3.4A). To build a comparison heatmap, a user specifies a new comparison name, selects a peptide filter if necessary, and specifies a comparison quantitation type. Four comparison types are already defined, *i.e.* single fraction/label-free, single fraction/MS1-Label, multiple fractions/label-free, and multiple fractions/MS1-Label. Single fraction experiment refers to a sample with only one fraction and multiple fractions experiment refers to a sample with many fractions such as collected in a MudPIT experiment. If peptides in the selected experiments are labeled with stable isotopes and quantified in MS scans such as SILAC, ICAT, GIST [23], or ICPL [24], an MS1-Label comparison type needs to be selected, otherwise a label-free type is chosen. For each additional LC/MS experiment to be incorporated into the heatmap, the user needs to specify a column number in the heatmap, a fraction number, and a replicate number (technical or biological replicates). If an exogenous standard peptide is spiked into each sample for normalization in label free experiments, each peptides' selected ion chromatogram (SIC, equivalent to an extracted ion chromatogram) peak area is normalized to the corresponding standard peptide SIC when building a comparison. To create a heatmap, FileMaker automatically collates a nonredundant list of unique peptides from different experiments but having the same peptide sequence, post-translational modifications and charge state. If there are replicate LC/MS datasets or multiple fractions, these are combined and the replicate quantitative data is averaged from the replicate experiments for each unique peptide. Essential peptide metadata such as SIC peak area, stable-isotope labeling ratio, mass error, Xcorr,

MOWSE score, logistic spectral score, phosphorylation site assignment, assigned peptide sequence, and protein name are transferred into the comparison table (Figure 3.4B).

PeptideDepot can generate heatmaps from the corresponding selected ion chromatogram peak areas from label free or stable-isotope labeled data using flexible equations defined within a custom function in FileMaker (Equation 3.1 and Equation 3.2). Two comparison heatmaps are generated using either label-free data or stable-isotope labeled quantitation. The label-free heatmap on the left (Figure 3.4B) represents the abundance of each peptide compared across all cellular states (optionally normalized to exogenous peptide standards if available). In this heatmap, a black color represents the average abundance for that unique peptide across all cellular states. A blue color indicates a selected ion chromatogram peak area that is less than the average and a yellow color indicates a peak area more than the average. The magnitude of the color change correlates with the magnitude of change in the underlying SIC peak areas. The stable-isotope labeling ratios heatmap is on the right (Figure 3.4B). In this representation, a black color represents a stable-isotope label ratio of 1:1 while a green color represents a ratio greater than 1 and a red color represents a ratio less than 1. For both types of heatmap, the magnitude of the CV for collated replicate LC/MS datasets is illustrated using an outline border surrounding the colored heatmap square. A black outline indicates a low CV amongst the replicate analyses and a bright green outline in label-free heatmap or bright red outline in the ratio heatmap indicates a high CV. Hovering the mouse pointer over any heatmap square reveals all underlying data corresponding to the generation of heatmap such as replicate peak areas, average peak areas, stable-isotope labeling ratio, and CV for replicate measurements (Figure 3.4D-E). Clicking any

heatmap square displays the selected peptides' SIC and profile MS spectral data (Figure 3.4C). This feature allows the manual adjustment of SIC peak boundaries and noise levels for all replicates separately. The user-defined peak parameters are imported back into FileMaker and the heatmap square is updated automatically and transparently for the end user.

$$\bar{x} = \sqrt{\max(x_{i,j}) \times \min(x_{i,j})}$$

$$f(x_{i,j}) = \begin{cases} 255 \times \dfrac{\left| Ln \dfrac{x_{i,j}}{\bar{x}} \right|}{Lnp_{\max}} & \dfrac{x_{i,j}}{\bar{x}} \le p_{\max} \\ 255 & \dfrac{x_{i,j}}{\bar{x}} > p_{\max} \end{cases}$$

$$color\ code = \begin{cases} \text{RGB}(255,255,255) & x_{i,j} < p_{\min} \\ \text{RGB}(0,0,f(x_{i,j})) & p_{\min} \le x_{i,j} < \bar{x} \\ \text{RGB}(0,0,0) & x_{i,j} = \bar{x} \\ \text{RGB}(f(x_{i,j}),f(x_{i,j}),0) & x_{i,j} > \bar{x} \end{cases}$$

$x_{i,j}$: SIC peak area of each timepoint (*i*) for each peptide (*j*)

max($x_{i,j}$): max peak area of peptide *j*

min($x_{i,j}$): min peak area of peptide *j*

$p_{\max}$: maximal fold change defined by the user

$p_{\min}$: minimal peak area defined by the user to be shown in the heatmap

RGB(*r,g,b*): function to generate a color based on three parameters. *r, g* and *b* representing the intensity of red, green and blue, ranging from 0 to 255.

Equation 3.1: Equation for color representations of relative peptide abundance in label-free heatmap

$$x_{i,j} = \begin{cases} r_{\max} & x_{i,j} > r_{\max} \\ x_{i,j} & r_{\max} \geq x_{i,j} \geq r_{\min} \\ r_{\min} & x_{i,j} < r_{\min} \end{cases}$$

$$color\ code = \begin{cases} \mathrm{RGB}(255 \times sqrt(\frac{r_{\min}}{x_{i,j}},2.5),0,0) & x_{i,j} < 1 \\ \mathrm{RGB}(0,0,0) & x_{i,j} = 1 \\ \mathrm{RGB}(0,255 \times sqrt(\frac{x_{i,j}}{r_{\max}},2.5),0) & x_{i,j} > 1 \end{cases}$$

$x_{i,j}$: selected SILAC ratio of each timepoint (*i*) for each peptide (*j*)

$r_{\max}$: maximal SILAC ratio change defined by the user

$r_{\min}$: minimal SILAC ratio change (reciprocal of the min number provided by the user)

sqrt(n,m): function to calculate *m*-th square root of number *n*

RGB(*r,g,b*): function to generate a color based on three parameters. *r, g* and *b* representing the intensity of red, green and blue, ranging from 0 to 255.

Equation 3.2: Equation for color representations of relative peptide abundance in SILAC heatmap

A variety of heatmap settings may be specified in a parameter pane available on the heatmap layout within FileMaker (Figure 3.4F). Minimal peak area thresholds can be set for both label-free and stable-isotope label ratio heatmaps to minimize the impact of low signal/noise data. The quantitative change thresholds (min and max fields) define the

minimum and maximum fold change and stable-isotope labeling ratios that are represented in the heatmap. Any changes below the minimal or above the maximal threshold in the heatmap are displayed as the same color as the minimum or maximum. The user can specify the CV value associated with the maximal heatmap outline color and specify whether to show the CV outline border.

3.3.2 Data Filtering and Extraction of Biological Significance of Proteomic Data

In addition to the quantitative comparison mode, identified peptides in a single LC/MS experiment can be explored in an informative layout to reveal their fundamental biological properties. The peptide listview (Figure 3.5B) has an extensive collection of predefined data filters. The currently implemented filters include Xcorr/charge state thresholding, MOWSE score thresholding, enzymatic cleavage type (for no enzyme type database searches), precursor ion mass error, manual validation status, protein name filtering, redundant peptide removal, logistic spectral score [25], thresholds for the number of phosphorylation sites and quality of the phosphorylation site localization (Ascore [26] thresholding), and maximal number of internal tryptic cleavage sites. Decoy database estimated false discovery rate (FDR) is dynamically calculated to provide the user with an estimate of the quality of the proteomic data and is updated as the user refines the filtering criteria.

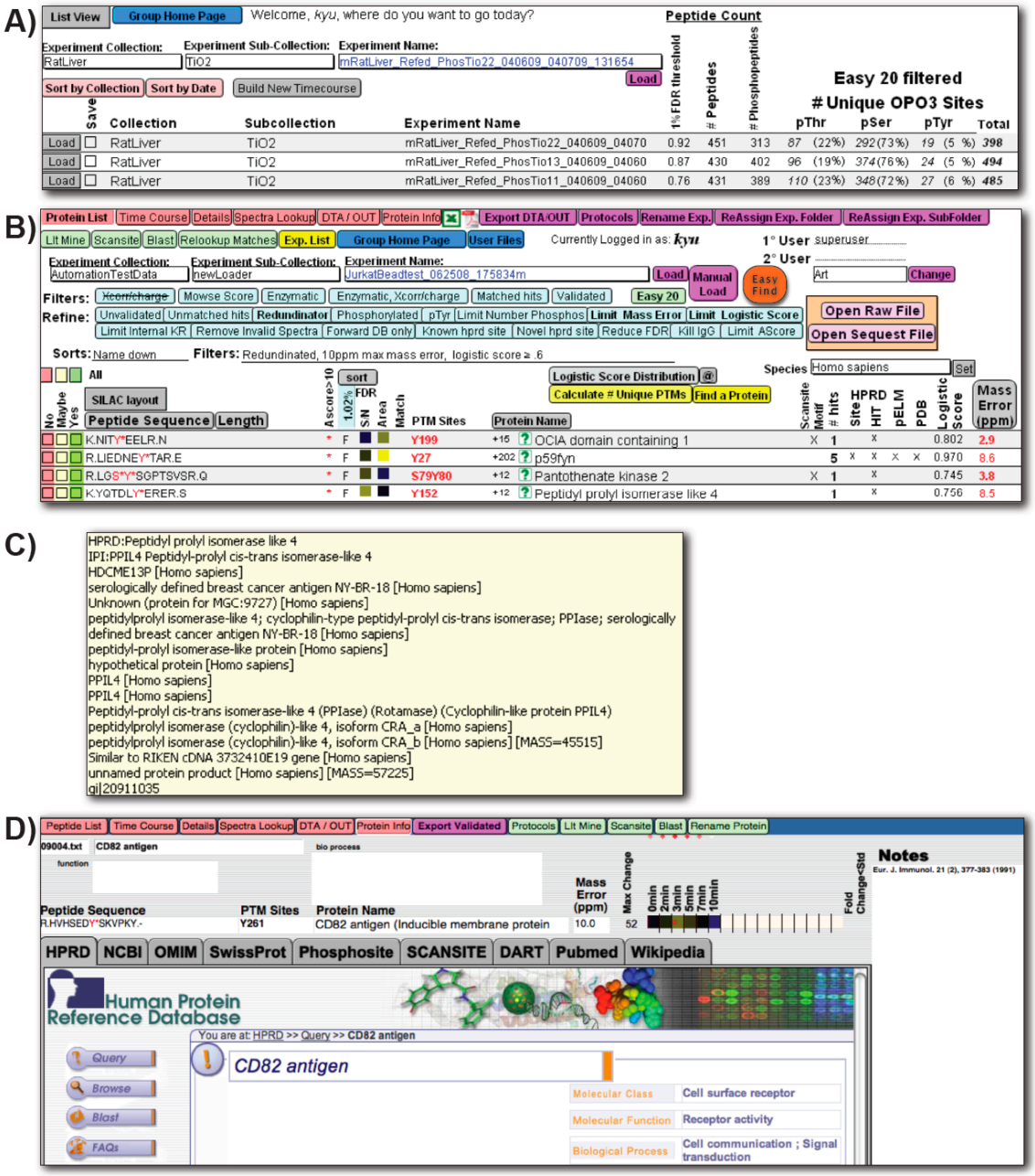


Figure 3.5: FileMaker generated graphical layouts for assimilation, comparison, and exploration of proteomic data and experimental metadata and collation of external protein information. A) Summary of experiments loaded into the PeptideDepot database after fully automated or manual post-acquisition analysis with user-customizable summary of numbers of peptides and types of peptide modifications observed; B) User-modifiable list of identified peptides from a single LC/MS experiment with an array of data filters and useful metadata for exploration of proteomic data; C) Moving mouse over the currently selected protein name shows a list of protein names matched to this peptide sequence collated from all internally indexed external genomic databases; D) external protein information, queried internally within FileMaker, displayed in a tabbed layout with quantitative proteomic data and a notepad for the user to document observations made during exploration. To explore the biological meaning of the proteomic data, a user may with a single click search for a certain peptide or protein among all independent protein information databases.

To enhance the utility of the database to biologists with limited proteomic knowledge, a database button provides a unified combination of filters defined by the proteomics lab to filter the data to a defined false-discovery rate. The proteomics lab can tailor these filters to the types of analyses and mass spectrometry equipment used in collecting the proteomic data in order to provide the most appropriate filters for distinct types of data. This function is useful to provide biologists with a unified collection of filters defined by the proteomic researcher that can be applied to the proteomic data to minimize inappropriate interpretation of the data.

Genomic database redundancy provides a significant challenge to efficient exploration of large proteomic datasets. Often a single peptide sequence will match many protein sequences contained within a given genomic database. Sometimes these peptide hits are actually entirely different proteins; more frequently, however, they are hits on redundant genomic database entries for the same peptide sequence with vastly different protein descriptors in the genomic database header. Current proteomic software tools associate the first protein hit achieved in the database search, regardless of the quality of the protein description. Although commercial software makes it possible to perform peptide lookup across the entire genomic database, the matches are not archived permanently with the data, and user-preferred associations between peptide sequence and meaningful protein names are not maintained between proteomic experiments. PeptideDepot ameliorates this deficiency by archiving user-selected peptide, protein name associations between proteomic experiments and caching associations between all peptides sequenced in an experiment with all genomic database proteins that contain that

peptide as the data is loaded manually or in an automated fashion with HTAPP (Figure 1).

To more rapidly understand the biological meaning of newly acquired proteomic data, a variety of genomic protein information databases are queried automatically. Currently, the protein information databases NCBInr [27], IPI [28], HPRD [29], Swiss-Prot [30], STRING [31], and Scansite [32] are queried by peptide sequence across locally cached copies of these databases as each proteomic experiment is loaded into PeptideDepot. All possible protein names associated with a given peptide sequence are cached automatically and displayed on demand (Figure 3.5C). A deep search through the multitude of redundant names for all assigned peptides across all protein information databases, allows the user to quickly ascertain if a certain protein was found in the experiment. (Figure 3.5B) This search capability overcomes the limitation of alternative protein naming across multiple protein information databases. By default, names from Swiss-Prot, IPI, or HPRD have priority over names from the redundant NCBInr database. The user can manually reassign the peptide to any matched genomic protein name hit with a single click. By maintaining a registry of user-preferred peptide sequence/genomic database associations, subsequent proteomic data is automatically associated first with user specified names that are most meaningful to the end-user, making proteomic data browsing much more efficient.

A separate PeptideDepot layout accessed by clicking the '?' icon next to protein name in the peptide list view allows users to explore protein annotation contained across all protein information databases (Figure 3.5D). Currently the peptide information view provides direct links to information from HPRD, NCBI Genbank, OMIM [33], ExPASy

Swiss-Prot, Phosphosite [34], the sequence motif analysis site Scansite, the protein domain analysis site CDART [35], the literature database PubMed, the online encyclopedia Wikipedia, and the protein-protein interaction database STRING. This tool eliminates the necessity to visit a multitude of web sites in an external web browser to gather protein information, thus making it possible to explore the significance of proteomic data directly within FileMaker through a tabbed interface where users insights into the meaning of proteomic data may be recorded directly within the database.

3.3.3 Peptide Validation

Estimation of false discovery rates using decoy database approach and statistical analysis of database sequence assignments is an important component of many proteomic workflows. PeptideDepot integrates a suite of existing tools to aid in this type of analysis [25]. The logistic spectral score is a new SEQUEST or Mascot rescoring algorithm that increases confidently assigned peptide assignments at a fixed FDR. FDR estimated using the decoy database approach is calculated for the experimental data being explored and is recalculated with each addition of a data filter or threshold (Figure 3.5B). Definition of an appropriate Xcorr, MOWSE score, or logistic spectral score threshold to achieve a user preferred, decoy database-estimated FDR is a tedious iterative process with current software tools. PeptideDepot increases user efficiency by calculating the score thresholds necessary to achieve a user's preferred FDR.

Manual MS/MS spectral validation is often an important but laborious part of many proteomic workflows. To expedite manual validation, we have integrated a recently developed database-integrated manual spectral annotation tool directly within a spectral

validation layout in PeptideDepot [25]. Using this existing Java software tool, a user may add any annotation to any MS/MS peak without leaving the database while all changes are recorded transparently to PeptideDepot. If the user revisits the manually annotated spectra at a later time, the user's manual annotations are retrieved transparently. If a user decides that the peptide assignment is valid, a single user click records the users validation and copies the peptide spectrum to a database of manually validated spectra.

3.3.4 Efficient Phosphoproteomic Data Analysis

PeptideDepot facilitates the analysis of post-translational modification (PTM) by calculating the phosphorylation site position and automatically determining whether the phosphorylation site has been described previously in the literature (Figure 3.5B). If the peptide contains a PTM present in the HPRD or phospho.ELM [36] databases, the user is automatically alerted to this fact and a single click reveals the journal article describing its discovery in PUBMED. If the peptide is contained in a protein within the protein structure database PDB [37], one user click reveals the three-dimensional protein structure in PyMOL (http://www.pymol.org).

3.3.5 Workflow Variability

PeptideDepot provides essential flexibility to support alternative proteomic workflows. The quantitative comparison component in PeptideDepot provides for the visual representation and automated interpretation of any MS-based quantitation method such as using a stable-isotope label or label free. In addition to Visual Basic quantitation software for Thermo .RAW files, PeptideDepot can display quantitative data calculated

from mzXML or mzData files using the previously described ProteinQuant [22] software. Quantitative data generated by any quantitation software are treated identically once they are imported into PeptideDepot.

End-users without any knowledge of programming can add or modify the result report forms simply through a drag-and-drop action using the FileMaker graphical WYSIWYG layout editor. New database fields and table relationships can be created within a user-friendly graphical database schema editor to add new data fields to the database or to define relationships between existing database fields. FileMaker also contains an intuitive scripting language that can automate arduous data manipulations or calculations including automation of calculations external to the database.

Unlimited numbers of external protein information databases can be directly integrated within PeptideDepot to meet any lab-specific needs for additional protein annotation. Data within PeptideDepot are accessible through the FileMaker client, a web browser, ODBC/JDBC, and PHP API, further expanding the ways in which a proteomic researcher can interact with the data.

3.3.6 Rapid Access to Proteomic Data

Currently our PeptideDepot database contains 16.5 million records of acquired peptide data and protein information from publicly available genomic databases. PeptideDepot provides rapid access to peptide data even in a table with millions of records. In speed trials, it took 2.7 seconds to load a dataset with 600 peptide records, and 7.3 seconds to load a 5,000-peptide dataset in PeptideDepot.

3.4 CONCLUDING REMARKS

The development of proteomic software lags behind the evolution of software for the analysis of large genomic datasets [38]. Proteomic researchers are reticent to distribute unpublished data beyond the borders of their labs, generating a great need for a flexible, integrated, open-source lab-based database for automated processing, interpretation, and visualization of proteomic data. We have presented a new approach to quantitative processing and visual analysis of proteomic data that directly integrates the data among different tools. Our system is designed to provide critical flexibility and easily implemented functionality to researchers in other proteomics labs, without the need for dedicated programmers. Our software integrates an array of essential functionalities including many unique features not currently found in any other publicly available integrated software package.

3.5 REFERENCE

1. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry.* J Mass Spectrom, 2001. **36**(8): p. 849-65.

2. Schwartz, J.C., M.W. Senko, and J.E. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* J Am Soc Mass Spectrom, 2002. **13**(6): p. 659-69.

3. Syka, J.E., et al., *Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications.* J Proteome Res, 2004. **3**(3): p. 621-6.

4. Yates, J.R., et al., *Performance of a linear ion trap-Orbitrap hybrid for peptide analysis.* Anal Chem, 2006. **78**(2): p. 493-500.

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

6. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.

7. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.

8. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.

9. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.

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

11. Salomon, A.R., et al., *Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry.* Proc Natl Acad Sci U S A, 2003. **100**(2): p. 443-8.

12. Taylor, C.F., *Minimum reporting requirements for proteomics: a MIAPE primer.* Proteomics, 2006. **6 Suppl 2**: p. 39-44.

13. Desiere, F., et al., *Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.* Genome Biol, 2005. **6**(1): p. R9.

14. Mathivanan, S., et al., *Human Proteinpedia enables sharing of human protein data.* Nat Biotechnol, 2008. **26**(2): p. 164-7.

15. Cottingham, K., *CPAS: a proteomics data management system for the masses.* J Proteome Res, 2006. **5**(1): p. 14.

16. Ulintz, P.J., et al., *4th Siena 2D Electrophoresis Meeting*. 2000, Siena, Italy.

17. Martens, L., et al., *PRIDE: the proteomics identifications database.* Proteomics, 2005. **5**(13): p. 3537-45.

18. Orchard, S., et al., *Current status of proteomic standards development.* Expert Rev Proteomics, 2004. **1**(2): p. 179-83.

19. Pedrioli, P.G., et al., *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol, 2004. **22**(11): p. 1459-66.

20. Deutsch, E., *mzML: a single, unifying data format for mass spectrometer output.* Proteomics, 2008. **8**(14): p. 2776-7.

21. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Mol Syst Biol, 2005. **1**: p. 2005 0017.

22. Mann, B., et al., *ProteinQuant Suite: a bundle of automated software tools for label-free quantitative proteomics.* Rapid Commun Mass Spectrom, 2008. **22**(23): p. 3823-34.

23. Chakraborty, A. and F.E. Regnier, *Global internal standard technology for comparative proteomics.* J Chromatogr A, 2002. **949**(1-2): p. 173-84.

24. Schmidt, A., J. Kellermann, and F. Lottspeich, *A novel strategy for quantitative proteomics using isotope-coded protein labels.* Proteomics, 2005. **5**(1): p. 4-15.

25. Yu, K., et al., *Integrated platform for manual and high-throughput statistical validation of tandem mass spectra.* Proteomics, 2009. **in press**.

26. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nat Biotechnol, 2006. **24**(10): p. 1285-92.

27. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2008. **36**(Database issue): p. D25-30.

28. Kersey, P.J., et al., *The International Protein Index: an integrated database for proteomics experiments.* Proteomics, 2004. **4**(7): p. 1985-8.

29. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

30. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.

31. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Res, 2005. **33**(Database issue): p. D433-7.

32. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.* Nucleic Acids Res, 2003. **31**(13): p. 3635-41.

33. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.

34. Hornbeck, P.V., et al., *PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation.* Proteomics, 2004. **4**(6): p. 1551-61.

35. Geer, L.Y., et al., *CDART: protein homology by domain architecture.* Genome Res, 2002. **12**(10): p. 1619-23.

36. Diella, F., et al., *Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.* BMC Bioinformatics, 2004. **5**: p. 79.

37. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

38. Topaloglou, T., *Informatics solutions for high-throughput proteomics.* Drug Discov Today, 2006. **11**(11-12): p. 509-16.

Chapter 4

TOOLS FOR MANUAL AND HIGH-THROUGHPUT STATISTICAL VALIDATION OF TANDEM MASS SPECTRA

4.1 INTRODUCTION

Liquid chromatography coupled with high-throughput mass spectrometry has become a powerful tool in proteomics. To interpret a tandem mass spectrum, peptide identification in LC/MS experiments follows one of two general approaches: (a) comparison of the obtained tandem mass spectrum to the theoretical spectrum corresponding to a database of sequences and (b) *de novo* construction of peptide sequences to match the obtained spectrum. Commonly used algorithms for the implementation of database searches include heuristic algorithms, e.g. SEQUEST [1], X!Tandem [2, 3] and Protein Prospector [4, 5]. Database search algorithms can also be based on probabilistic scoring such as MASCOT [6]. Algorithms for the implementation of *de novo* sequencing include those implemented in the software Lutefisk [7], and PEAKS [8]. Methods for peptide identification which combine aspects of database searching and *de novo* sequencing have also been proposed such as GutenTag [9] and InsPecT [10].

The identification of peptides via database searches typically results in a large number of candidate peptides. If a final assignment is determined by picking the "best" match from the search algorithm output, a substantial portion of the final selections may be incorrect due to the poor ionization and fragmentation efficiency of peptides, especially phosphopeptides. This necessitates the validation of thousands of spectra per experiment. A number of statistical methods have been developed to automate validation of large-scale datasets [5, 11-19]. These approaches combine key output from identification algorithms with other available information and also include statistical modeling in order to make reliable predictions of whether a "best" match is correct. A number of other

recent proposals in the literature use advanced tools from discriminant and cluster analysis to improve on the performance of peptide identification algorithms and to develop statistical measures of performance [9, 11, 16, 18, 20-23]. Recent work has applied statistics to the validation of phosphorylation sites from CAD-MS/MS data using statistical multiple testing and a support vector machine analysis [11], Bayesian network scoring [24], or target-decoy approach with a probability based phosphorylation site localization score [13, 25].

Although some researchers prefer to train models on manually validated datasets, others prefer the use of single protein digests or thresholding on spectral parameters while optimizing decoy database tests to train their models. Unfortunately, flexible training of existing algorithms to user specified validated datasets is not a feature that is currently directly supported in existing software tools, reducing their flexibility to alternative proteomic workflows. For example, MS/MS spectra from phopshorylated peptides have different spectral characteristics such as the abundant neutral loss of phosphate when compared to unphosphorylated peptides. A model trained on phosphorylated datasets may more closely capture the characteristics unique to correctly assigned phopshopeptide spectra, compared to a model trained on datasets lacking phosphorylated peptides. Furthermore, the validation philosophy employed to generate the training dataset may also impact the subsequent performance of a model when applied to a newly acquired dataset.

On the other hand, statistical validation alone is not enough to reach a determinate conclusion in many cases, given that no algorithm can predict with 100% accuracy all the spectra that are correctly assigned. Assigned peptides with significant biological

significance still require manual validation of both sequence and sites of covalent modifications to minimize ambiguity [26-28]. Manual spectral validation will continue to be an important part of any proteomic workflow. Tools increasing the efficiency of this arduous task are critical.

Commercial software for proteomic analysis such as Bioworks (Thermo Scientific) or Mascot [6] provides only static representations of assigned spectra, with no capability for user-driven manual annotation. One newly developed software tool, CHOMPER, enables highlighting of fragment peaks that are associated with certain user selected amino acids from spectra loaded manually from dta and out files [29]. CHOMPER also adds the capability for users to store decisions of overall spectral quality electronically.

An ideal electronic spectral annotation tool should meet the following requirements: a) automatically calculate theoretical fragment ion masses including neutral losses from the precursor and fragment ions, b) allow users to add any annotations to an MS/MS spectrum and save them for the future reference within a relational database, and c) tight integration within an automated proteomic pipeline. The goal of manual validation is the exhaustive assignment of all fragment ions observed in a spectrum. Often proteomic end-users are forced to mentally calculate theoretical neutral loss fragment ion masses with existing tools. Furthermore, current software tools designed to assist in the manual annotation of spectra are not integrated within lab-based relational databases. If the same peptide is observed in another dataset potentially collected by a different investigator, the user of existing software is not able to compare it to previously manually annotated spectra associated with that sequence, increasing the chances of redundant manual validation and decreasing overall laboratory efficiency.

We have incorporated the improvements and addressed the limitations of existing software for MS/MS validation by developing a comprehensive validation solution. This solution includes a program for generation of statistical models based on user validated datasets, integration of these user created models within automated proteomic workflows, and a unique visualization and annotation tool for manual spectral validation called SpecNote. These programs are fully integrated into the HTAPP platform.

4.2 MATERIALS AND METHODS

4.2.1 Software Architecture

We have designed the HTAPP where LC/MS data acquisition and post acquisition analysis is fully automated (Figure 4.1). This custom-made software controls multi-dimensional LC separations of peptides (with Immobilized Metal Affinity Chromatography capability), LC/MS data acquisition, and post acquisition SEQUEST search (version 27; Thermo Scientific), peptide quantitation, decoy database analysis, spectral validation, phosphosite localization using Ascore [11], and loading data into a lab-based relational database called PeptideDepot created in FileMaker (version 9.0.3; FileMaker Inc.) with live connections to data warehoused in a MySQL database (version 5.1.16-beta-nt; MySQL Inc.). In this manuscript we describe the peptide validation component of HTAPP. Within this validation component, the logistic-model-based validation program is launched automatically, without user intervention. A spectral score for each peptide is calculated by an R (version 2.4.1; GNU project) program, which indicates the probability that the SEQUEST generated sequence is correctly assigned

(1=most likely, 0=least likely). The estimation of false discovery rates (FDR) is accomplished by the decoy database approach and evaluated without user intervention. A figure illustrating the distribution of spectral score or XCorr versus decoy database search direction is dynamically generated within PeptideDepot from data stored in a custom MySQL table that represents the peptides currently viewed by the proteomic researcher. This data is represented within a custom webpage viewed through a FileMaker web portal and generated by a PHP script (ver. 5.2.1, http://www.php.net) hosted on Apache web server (ver. 2.2.4, the Apache Software Foundation), along with a table that lists the yields at a certain FDR.

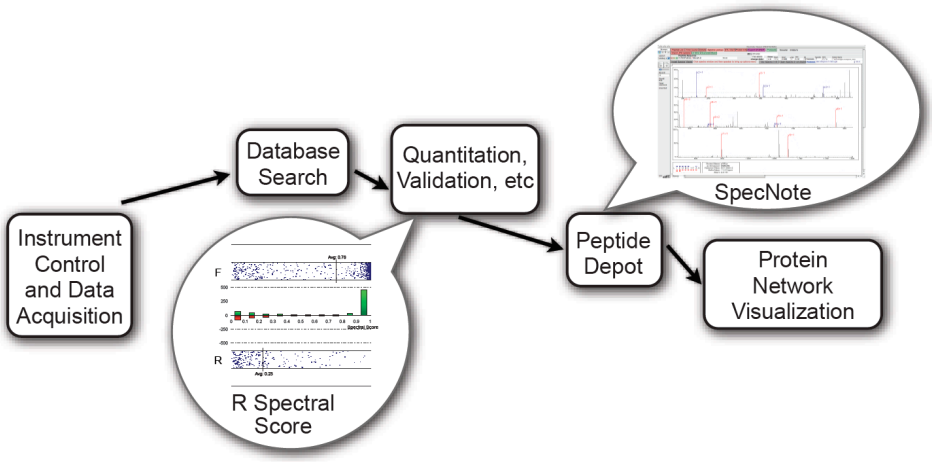


Figure 4.1: Schematic representation of how the statistical validation and manual validation software components fit into the HTAPP proteomic pipeline. Balloons indicate the software described in this chapter.

The manual annotation software component SpecNote, which is built based on publicly available Java libraries, including interfascia.jar, pde.jar, pdf.jar, itext.jar, jogl.jar, core.jar (all from Processing distribution 0125; http://processing.org/) and mysql-connector-Java.jar (version 5.0.5; MySQL Inc.), and compiled as a Java Applet (version

1.6; Sun Microsystems) that runs in any web browser, is embedded in a web portal within PeptideDepot. A user may navigate LC/MS experiment within PeptideDeopt, select any peptide and electronically annotate and validate the SEQUEST assigned MS/MS spectra within SpecNote. The user annotations are transparently stored in the MySQL relational database component of PeptideDepot, accessible to other users who discover the same peptide in another proteomic experiment.

SpecNote flexibly integrates within existing user workflows that may be independent of PeptideDepot. If a user already has software for selecting a certain peptide from a LC/MS experiment, SpecNote could be utilized through a webpage. If a user's manual validation workflow requires additional data to be represented on the spectra, a user may alter the source code of SpecNote to customize the GUI or to utilize alternative data sources such as a different relational database.

4.2.2 Experimental Datasets

Four datasets were chosen to train and evaluate the newly created logistic spectral model. The datasets represented a variety of typical proteomic data types including simple and complex mixtures of either phosphorylated or unphosphorylated peptides acquired on an LTQ mass spectrometer. The samples were 1) *Mast cell phosphopeptides (MCP5)*, 2) *Pervanadate stimulated T cell phosphopeptides (PVIP)*, 3) *18 protein ISB standard protein mix* [30] *(18Mix),* and 4) *Bovine serum albumin peptides (BSA)*. An additional dataset, *NIH3T3 phosphopeptides* (*3T3*) was prepared to test the performance of the logistic model on an LTQ/FTICR dataset.

Data Set 1 - *Mast cell phosphopeptides (MCP5)*. Transformed primary bone marrow-derived mast cells (MCP5) were stimulated as described previously. [26] Stimulated cells (~$5x10^8$) were lysed for 10 min with rotation at 4°C in 50 mL of lysis buffer consisting of 20 µg/mL aprotinin, 20 µg/mL leupeptin, 50 mM Tris pH 7.5, 100 mM NaCl, 1% Triton X-100, 10% glycerol, 1 mM Perfabloc, 2 mM Na_3VO_4, 10 mM β-glycerophosphate and 1 mM EDTA (all from Sigma, St. Louis, MO).

Lysates were centrifuged at 12,000x g for 20 minutes at 4°C. Fifty pmol of LIEDAEpYTAK per ~$5x10^8$ cell equivalents (c.eq.) and anti-phosphotyrosine agarose (clone PT66, Sigma; 200 µL resin/~$5x10^8$ cell equivalents) was added to the supernatant for 4 hr at 4° C with rotation. Beads were washed once with 50 mL lysis buffer and once with 50 mL of 20 mM Tris buffer, pH 7.4, 120 mM NaCl. Proteins were recovered from the beads with 100 mM NH_4HCO_3 buffer pH 8.3 containing 8 M urea for five minutes at 96° C and the supernatant was filtered using PVDF membranes with a pore size of 0.22 µm (Millipore Inc., Bedford, MA). The mixture was diluted with an equal volume of water and proteins were digested overnight with 5 µg of modified trypsin (Promega, Madison, WI) at 37° C.

Tryptic peptides were converted to methyl esters with deuteromethanol so that every Asp, Glu and peptide c termini contained an additional mass of 17 Da. Phosphopeptides were enriched with an automated desalt/immobilized metal affinity chromatography (IMAC)/nano-liquid chromatography/electrospray ionization mass spectrometry platform as described previously. [31] Peptides where eluted with a 30 minute 0-70% solvent B reversed-phase gradient through an analytical column with integrated 4 µm electrospray tip into an LTQ mass spectrometer with 30 nl/min peak parking (solvent A - 0.1 M acetic

acid ddH2O, solvent B - 0.1 M acetic acid acetonitrile). [31] Peptides were analyzed by data dependent tandem mass spectrometry (MS/MS) experiments using collision-induced dissociation (Xcalibur 1.4 parameters designated 35% collision energy, 3 Da isolation window, top 5 data dependent, repeat count of one, and a dynamic exclusion time of 1.5 min, IT-MS AGC of 30,000, and IT-MS/MS AGC of 10,000). MS/MS spectra were assigned to peptide sequences from the NCBI non-redundant protein database sliced in Bioworks 3.1 for human proteins and searched with the SEQUEST algorithm and X!Tandem. Search paramaters designated a static modification of +17.0342 Da on Asp, Glu and the c-terminus (deuteromethyl esters) and variable modifications of +79.9663 Da on Ser, Thr, and Tyr (phosphorylation).

Data Set 2 - *Pervanadate stimulated T cell phosphopeptides (PVIP)*. Cell culture was performed as described. [32] Briefly, Jurkat cells (clone E6-1) were grown in RPMI 1640 medium with 10% fetal bovine serum, 2 mM L-glutamine, 100 ug/ml streptomycin sulfate, and 100 U/ml penicillin G (all from Sigma) in a 5.0% CO_2 incubator. Cells were treated with pervanadate to inhibit tyrosine phosphatases and elevate levels of phosphotyrosine as described previously. [33] Cells ($1x10^9$) were washed with RPMI lacking FBS at 4°C then lysed for 20 min with rotation at 4°C in 25 mL of lysis buffer. Cellular lysates were purified and analyzed as in Data Set 1 with the exception that the amount of cell equivalents analyzed was reduced to $1x10^8$ as compared with $5x10^8$ for Data Set 1. MS/MS spectra were assigned to peptide sequences from the NCBI non-redundant protein database sliced in Bioworks 3.1 for human proteins and searched with the SEQUEST algorithm and X!Tandem. Search paramaters designated a static

modification of +17.0342 Da on Asp, Glu and the c-terminus (deuteromethyl esters) and variable modifications of +79.9663 Da on Ser, Thr, and Tyr (phosphorylation).

Data Set 3 - *Standard protein mixture (18Mix)*. The publicly available raw data files acquired on LTQ instrument and the validated peptide assignments were downloaded from http://regis-web.systemsbiology.net/PublicDatasets/ [30]. MS/MS spectra were assigned to peptide sequences from the NCBI non-redundant protein database sliced in Bioworks 3.1 for *H. influenzae* and searched with the SEQUEST algorithm. SEQUEST search paramaters designated a static modification of +57.0215 Da on Cys (alkylation).

Data Set 4 - *Bovine serum albumin peptides (BSA)*. Bovine serum albumin (BSA; Sigma) was reconstituted in 8M urea, 100 mM ammonium bicarbonate, pH 8.0. Tris-carboxyethyl phosphine was added to 10 mM and the mixture allowed to stand at room temperature for 10 minutes. Iodoacetamide was then added to a final concentration of 20 mM, and the mixture was incubated at 22 ˚C for 45 min. in the dark. The mixture was then diluted with an equal volume of 100 mM ammonium bicarbonate, pH 8.0. Modified trypsin (Promega) was added, and the mixture incubated for 8 hr at 37 °C. BSA peptides were enriched with our automated system as in Data Set 1 except the desalt and IMAC separations were skipped. [31] Peptides where eluted with a 30 minute 0-70% solvent B reversed-phase gradient through an analytical column with integrated 4 µm electrospray tip into an LTQ mass spectrometer with no peak parking (solvent A - 0.1 M acetic acid, solvent B - 0.1 M acetic acid acetonitrile). [31] Data acquisition parameters were identical to Data Set 1. MS/MS spectra were assigned to peptide sequences from the bovine NCBI non-redundant protein database with the SEQUEST algorithm and

X!Tandem. Search parameters designated a static modification of +57.0215 Da on Cys (alkylation).

Data Set 5 – *Insulin stimulated 3T3 phosphopeptides (3T3)*. Cell culture was performed as described [34]. Briefly, 3T3 cells transfected with IRS-1 were incubated with DMEM supplemented with 10% FBS, 2 mM L-glutamine, and 400 ug/ml G418 (all from Sigma except L-Glu which from Invitrogen) in a 5.0% CO_2 incubator. After reaching 95% confluence, Cells were starved in DMEM with 0.1% BSA for 24 hours. On the next day, cells were stimulated with 100 nM insulin for 5 minutes and then immediately lysed in cold lysis buffer consisting of 8 M urea, 100 mM NH_4HCO_3 and 1 mM NaV_3O_4 for 20 minutes. The lysates were reduced, alkylated, digested, and desalted as described [26]. Purified peptides were then fractionated with a homemade SCX column (500 μm x 15 cm PEEK tubing (Upchurch, Oak Harbor, WA) packed with 5 μm PolyLC SCX resin (The Nest Group, Southborough, MA)). Phosphopeptides in each fraction were enriched by TiO_2 as described [25] and analyzed on LC/MS as in Data Set 3. MS/MS spectra were assigned to peptide sequences from the human NCBI non-redundant protein database with the SEQUEST algorithm and X!Tandem. Search parameters designated a static modification of +57.0215 Da on Cys (alkylation) and variable modifications of +79.9663 Da on Ser, Thr, and Tyr (phosphorylation).

Using the SEQUEST algorithm, tandem mass spectra were assigned to peptide sequences from species-specific NCBI non-redundant protein databases. The forward NCBI databases were reversed and appended to the forward database to estimate the false discovery rate [35]. SEQUEST search parameters varied depending on the dataset as described in supplemental materials. For all datasets, search parameters designated

tryptic enzymatic cleavage. SEQUEST results were thresholded on XCorr (+1>1.5, +2>2, +3>2.5). For comparison, the same datasets were searched using X!Tandem database algorithm (version 2008.12.01.1) with the identical search parameters and protein databases used in SEQUEST searching. X!Tandem results were thresholded on E-value (≤1.0) for LTQ data, or on precursor mass error (≤20 ppm) for LTQ-FTICR data.

The MCP5 dataset consisted of 1114 spectra, with 630 valid and 484 invalid spectral assignments determined by manual validation. The PVIP set consisted of 619 spectra with 193 manually validated as correct assignments; the remaining 426 were determined to be incorrect. The BSA set consisted of manually validated 605 spectra with 303 correct and 302 incorrect. The 18Mix set consisted of 25856 spectra with 14568 assigned correctly to the 18 proteins known to be in this sample while 11288 were assigned as incorrect because they were not amongst the 18 known proteins. A subset of the 18Mix spectra was randomly selected for the model training and evaluation, including 913 valid and 687 invalid assignments.

4.2.3 Criteria for Manual Validation of Spectra

Spectra were passed through intensive manual validation to ascertain whether SEQUEST assigned sequences were consistent with MS/MS spectra for all datasets except 18Mix. Our requirements were identical to previously described manual validation metrics [36] with the additional requirements that 1) threonine and serine phosphorylated peptides should contain an abundant neutral loss of phosphate from the precursor ion (M-80/z Da or M-98/z Da), 2) all abundant peaks should be assigned to either a b or y ion or a neutral loss of phosphate, water, or ammonia from a b, y, or

precursor ion, 3) only monoisotopic peaks are assigned, and 4) at most two internal cleavage sites were allowed for samples digested with trypsin and peptides containing any internal cleavage sites were scrutinized more closely.

4.2.4 Statistical Methods for Spectral Validation

Logistic regression was used to develop statistical models for peptide validation, with the response variable indicating whether the peptide identification is valid or invalid. Three manually validated datasets (MCP5, PVIP, BSA) along with another dataset validated by matching to a known mixture of 18 proteins (18Mix) were used as training sets. For each spectrum, a number of predictor variables believed to mimic manual validation criteria were calculated and used to fit a logistic regression model. There were four groups of predictors that in total constituted 34 variables as described (Table 4.1).

(a) Variables obtained directly from SEQUEST output (variables 1-7): These variables are all computed by SEQUEST and include $XCorr_1$, charge state, MH $mass_1$, sp_1, and the ions $ratio_1$ (separated into numerator, denominator, and computed ratio multiplied by 100). These variables were taken from the peptide SEQUEST identified as the most likely match, the first peptide in the "OUT" file.

(b) Variables computed from SEQUEST output (variables 8 - 16): These variables included $Xcorr' = \left(\frac{\ln(Xcorr_1)}{\ln(2 * Charge * NumberAminoAcids)} \right) * 10$ (as calculated with PeptideProphet [12]), as well as $\Delta C_2 = \frac{Xcorr_1 - Xcorr_2}{Xcorr_1}$

The subscript 2 in all variables denotes the second highest-ranking peptide in the SEQUEST OUT file that differs from the first peptide in peptide sequence and *not* in the location of the phosphorylation site.

$\Delta Mass_1$ and $\Delta Mass_2$ compare the theoretical mass of the assigned peptide sequence with the observed experimental mass. The variables in this group also include the number of amino acids in the first peptide, the number of internal enzymatic cleavage sites in the first peptide, and the numbers of phosphorylated S, T, and Y amino acids in the first "OUT" file peptide.

(c) Variables computed directly from the spectra (variables 17 - 18): These variables are used as a general measure of noise in the spectra. After normalizing peak intensities so that the largest intensity is 1, the mean and median are computed.

(d) Variables computed by comparing the spectra and SEQUEST output (variables 19 – 34): Variables in this category ascertain the degree to which the spectrum corresponds to the top SEQUEST reported peptide using insight from expert manual validation. First, the overall spectrum quality is evaluated by *fancymean* that calculates the relative abundance of those important peaks comparing to noise peaks. The spectrum is studied for peaks that correspond to neutral losses from the precursor mass. The variable *phosphoscore* is a ranked average of peak heights for peaks that correspond to a loss of a phosphate from the precursor mass if the peptide SEQUEST has selected is phosphorylated. A ranked average of peaks corresponding to other neutral losses from the precursor mass, such as ammonia and water, is determined in the variable *sumscore.*

Next the spectrum is examined for peaks unassigned to a y or b ion or neutral loss peak. An experimental MS/MS peak was matched with a theoretical assignment only if

within 0.5 Da of the peak. Although we used 0.5 Da, this mass error is a user specified tolerance. The assigned peaks were classified into three groups; *Group 1* - b and y ions, *Group 2* - neutral loss of a single ammonia, water, or phosphate from a b ion, y ion, or precursor ion, and *Group 3* - neutral losses of multiple ammonia, water, or phosphate from a b ion, y ion, or precursor ion.

Comparisons are made between the sum and average intensities of the assigned peaks to the sum and average intensities of the *important* peaks. These variables will be low for spectra where many *important* peaks remain unassigned.

Finally, general sequence coverage of the peptide is studied by computing the fraction of *Group 1-3* theoretical ions assigned to experimental spectral peaks.

The "SEQUEST" model was developed using only the variables in group (a); the "SEQUEST Plus" model uses variables from groups (a) and (b); and the "Spectral" model was determined by using stepwise reduction on the variables in all four groups. Unlike the SEQUEST and SEQUEST Plus models, the final variables retained in the Spectral model depended on the training dataset. The stepwise reduction began with a full list of variables, sequentially removing each variable and comparing the distribution of model predicted likelihoods to determine if the performance of the model was significantly changed (p-Value<0.05). In particular, the remaining number of variables retained after stepwise model reduction was 13 for MCP5, 8 for PVIP, 9 for BSA, and 13 for 18Mix. The resulting models were applied to the validated datasets and ROC analysis was used to assess their predictive performance. The ROC curves were summarized and compared via their corresponding areas under the curve (AUC). All computations were performed

using either STATA software (ver.10, StataCorp LP) or software written in R (version 2.4.1; GNU project). In order to compare the effectiveness of the spectral model at a low FDR estimated by decoy database, the number of peptide hits found in the proteomic datasets thresholded by the Spectral score, XCorr or X!Tandem E-value to achieve an overall 1% FDR were counted and summarized into a table.

Table 4.1: Full list of variables used in model computing

	Variable Description	Variable Name in Model
(a) Variables obtained directly from SEQUEST output:		
1	Charge state of the precursor ion	charge
2	Xcorr value of the top SEQUEST hit	xcorr1
3	MHMass = Mass of the top SEQUEST hit	mhmass
4	Ions numertor for top SEQUEST hit	ionsnum
5	Ions denominator for top SEQUEST hit	ionsden
6	The ratio of ions numerator to ions denominator	ionsratio
7	sp for top SEQUEST hit	sp
(b) Variables computed from SEQUEST output:		
8	The number of amino acids in the top SEQUEST hit	aa
9	Xcorr': xcorr1 normalized by its peptide length	xcorrp
10	ΔCn_2 : difference in xcorr1 and the xcorr value of the second highest ranking peptide reported by SEQUEST that differs from the first in peptide sequence and not in location of phosphorylation site.	dc2
11	Number of internal K and R amino acids in top SEQUEST hit	kr
12	Relative difference in mass of top SEQUEST hit and experimental mass	dmass1
13	Relative difference in mass of next sequence different SEQUEST hit and experimental mass	dmass2
14	Number of phosphorylated S amino acids	ps
15	Number of phosphorylated T amino acids	pt
16	Number of phosphorylated Y amino acids	py
(c) Variables computed directly from only the spectrum:		
17	Median of normalized peak intensities	median
18	Mean of normalized peak intensities	mean

	(Peak intensities are normalized so that the maximum peak intensity is 1.)	

(d) Variables computed by comparing the spectrum and SEQUEST output:

Accounting for Neutral Loss of Water, Phosphate, and Ammonia

19	Ranked weighting of the number of peaks corresponding to neutral loss of phosphate off the precursor mass. (*)	phosphoscore
20	Sum of intensities for any peak in the spectra corresponding to some neutral loss (**).	sumscore

Accounting for Noise from Peak Intensities

21	The total number of peaks in the spectrum divided by aa	number
22	Relative abundance of mean peak intensity of all peaks over the mean intensity of low abundance peaks	fancymean
23	Number of ASPs (****) divide by number of IMPs	noa
24	Number of *Group 1* assigned peaks divided number of ASPs	nda
25	Number of *Group 1-2* assigned peaks divided by ASPs	nsa
26	Sum of intensities ASPs divided by sum of intensities of IMPs	toa
27	Sum of intensities of *Group 1* assigned peaks divided by sum of ASPs	Tda
28	Sum of intensities of *Group 1-2* assigned peaks divided by sum of ASPs	tsa
	Accounting for Sequence Coverage	
29	Fraction of unassigned y and b-ions	percunass
30	Fraction of unassigned or *Group 3* assigned y and b-ions	percweakass
31	Fraction of unassigned, *Group 2-3* assigned y and b-ions	percondirass
32	Fraction of y and b-ions with more than one peak assignment of any kind	onehit
33	Fraction of y and b-ions with more than one *Group 1-2* hit	onestronghit
34	Fraction of y and b-ions with more than one *Group 1* hit	onedirecthit

(*) If more than one phosphorylation site in the peptide we account for loss of single or multiple phosphates as well as loss of phosphate and water or ammonia.

In the case of a singly phosphorylated peptide, phosphoscore is a number between 0 and 4.

+2 if the neutral loss of phosphate is directly assigned,
+1 if we assign the loss of phosphate and water,
+1 if we assign the loss of phosphate and ammonia

(In the event of multiple phosphorylation sites, we still look for single losses first)
(**) We consider loss of phosphate, water, ammonia and any combination of these.
(***) IMP = The *important* peaks; The top N intensive peaks in spectrum. N=peptide length
(****) ASP = The assigned *Group 1-3* peaks.

4.3 RESULTS AND DISCUSSION

4.3.1 Statisitical Validation

Each of three logistic regression models was applied to all four validated datasets. The AUC was computed and compared to those of the single SEQUEST variable XCorr (ΔAUC) (Table 4.2). All three logistic regression models: SEQUEST, SEQUEST Plus, and Spectral, performed statistically better than XCorr in most cases (ΔAUC p-value < .05). Amongst our models the spectral model performed the best compared to XCorr with a statistically significant Δ AUC for all but one case. The SEQUEST model performed the poorest with 3 out of 16 cases of no significant change between XCorr and SEQUEST model and one case where XCorr outperformed the SEQUEST model. The spectral model trained on MCP5 was selected for further analysis because it resulted in the highest ΔAUC of all models when cross-applied to the other datasets.

Table 4.2: AUC of SEQUEST, SEQUEST Plus, and Spectral models trained on and applied to all datasets.

	Applied To:	BSA		PVIP		MCP5		Standard Mix	
Training Set		AUC	ΔAUC p-Value	AUC	ΔAUC p-Value	AUC	ΔAUC p-Value	AUC	ΔAUC p-Value
BSA	Sequest	0.761	0.070 <0.001	0.824	-0.006 0.733	0.896	0.136 <0.001	0.869	0.044 <0.001
	Sequest Plus	0.812	0.121 <0.001	0.804	-0.026 0.241	0.921	0.161 <0.001	0.885	0.060 <0.001
	Spectral	0.903	0.212 <0.001	0.837	0.007 0.739	0.908	0.148 <0.001	0.874	0.049 <0.001
PVIP	Sequest	0.693	0.002 0.841	0.900	0.07 <0.001	0.858	0.098 <0.001	0.867	0.042 <0.001
	Sequest Plus	0.716	0.025 0.012	0.936	0.106 <0.001	0.896	0.136 <0.001	0.862	0.037 <0.001
	Spectral	0.852	0.161 <0.001	0.947	0.117 <0.001	0.920	0.160 <0.001	0.875	0.050 <0.001
MCP5	Sequest	0.762	0.071 <0.001	0.862	0.032 0.029	0.920	0.160 <0.001	0.875	0.050 <0.001
	Sequest Plus	0.782	0.091 <0.001	0.890	0.060 <0.001	0.961	0.201 <0.001	0.890	0.065 <0.001
	Spectral	0.849	0.158 <0.001	0.897	0.067 <0.001	0.970	0.210 <0.001	0.892	0.067 <0.001
Standard Mix	Sequest	0.662	-0.029 0.192	0.744	-0.086 <0.001	0.800	0.004 0.006	0.891	0.066 <0.001
	Sequest Plus	0.745	0.054 0.013	0.875	0.045 0.023	0.932	0.172 <0.001	0.909	0.084 <0.001
	Spectral	0.796	0.105 <0.001	0.883	0.053 0.011	0.944	0.184 <0.001	0.914	0.089 <0.001
	XCorr	0.691		0.830		0.760		0.825	

As a comparison, the difference in AUC from XCorr (ΔAUC) is shown with the p-value below. Grey shading indicates a model trained and applied to the same dataset. The SEQUEST, SEQUEST Plus, and Spectral models outperform XCorr in all but four cases. The Spectral model performs well across all training and application datasets and has comparatively larger ΔAUC values among all models, with only one case that the Spectral model is not significantly better than XCorr (p-values < 0.05 adjusted for multiple comparisons).

To compare the performance of the spectral model to XCorr, we also examined the effect of the spectral score upon the distribution of forward and reversed database hits. The use of decoy estimated FDR provides a universal metric that allows comparison of the performance of user generated logistic models with other validation approaches. The distribution between the spectral score or XCorr and decoy database direction was

examined for both LTQ and LTQ/FTICR data (Figure 4.2; Table 4.3). This view is useful for selection of spectral score thresholds for user preferred FDR and is incorporated into our automated proteomic workflow using a dynamic PHP script (Figure 4.3D-E). With both LTQ and LTQ/FTICR data, the forward and reversed populations using the spectral score were significantly separated when compared to XCorr or XTandem E-value (Figure 4.2). This increased separation has the impact of increasing peptide yield at a user selected FDR. For instance, to reduce the FDR of 3T3 dataset to 1%, thresholding on spectral score retains 455 peptides out of a total of 959 peptides, comparing to 122 peptides by XCorr and 300 peptides by E-value. For the datasets mentioned in this paper, our logistic spectral model outperformed both XCorr (242% more peptides identified on average) and the X!Tandem E-Value (87% more peptides identified on average) at a 1% false discovery rate estimated by decoy database approach.

Table 4.3: The spectral model provides a substantial yield increase of confident peptide assignments when applied to a range of different proteomic datasets.

Datasets	Average Spectral Score			Average XCorr			Average –log(E-value)		
	Forward hits	Reversed hits	#hits at 1% FDR	Forward hits	Reversed hits	#hits at 1% FDR	Forward hits	Reversed hits	#hits at 1% FDR
BSA	0.81	0.47	58	3.2	2.7	41	1.42	0.36	82
PVIP	0.44	0.19	212	2.9	2.8	48	1.15	0.38	80
MCP5	0.47	0.19	637	3.0	2.7	154	0.91	0.36	246
3T3	0.76	0.23	455	2.9	2.5	122	1.16	-0.9	300

Spectral score is computed using the Spectral model trained on MCP5 dataset. E-value is computed by X!Tandem and represented as –log(E-value). For each dataset, thresholded by the Spectral score, XCorr or X!Tandem E-value, number of peptide hits at 1% FDR estimated by decoy database search is calculated. The spectral model outperformed both SEQUEST XCorr (242% more peptides identified on average) and X!Tandem (87% more peptides identified on average)

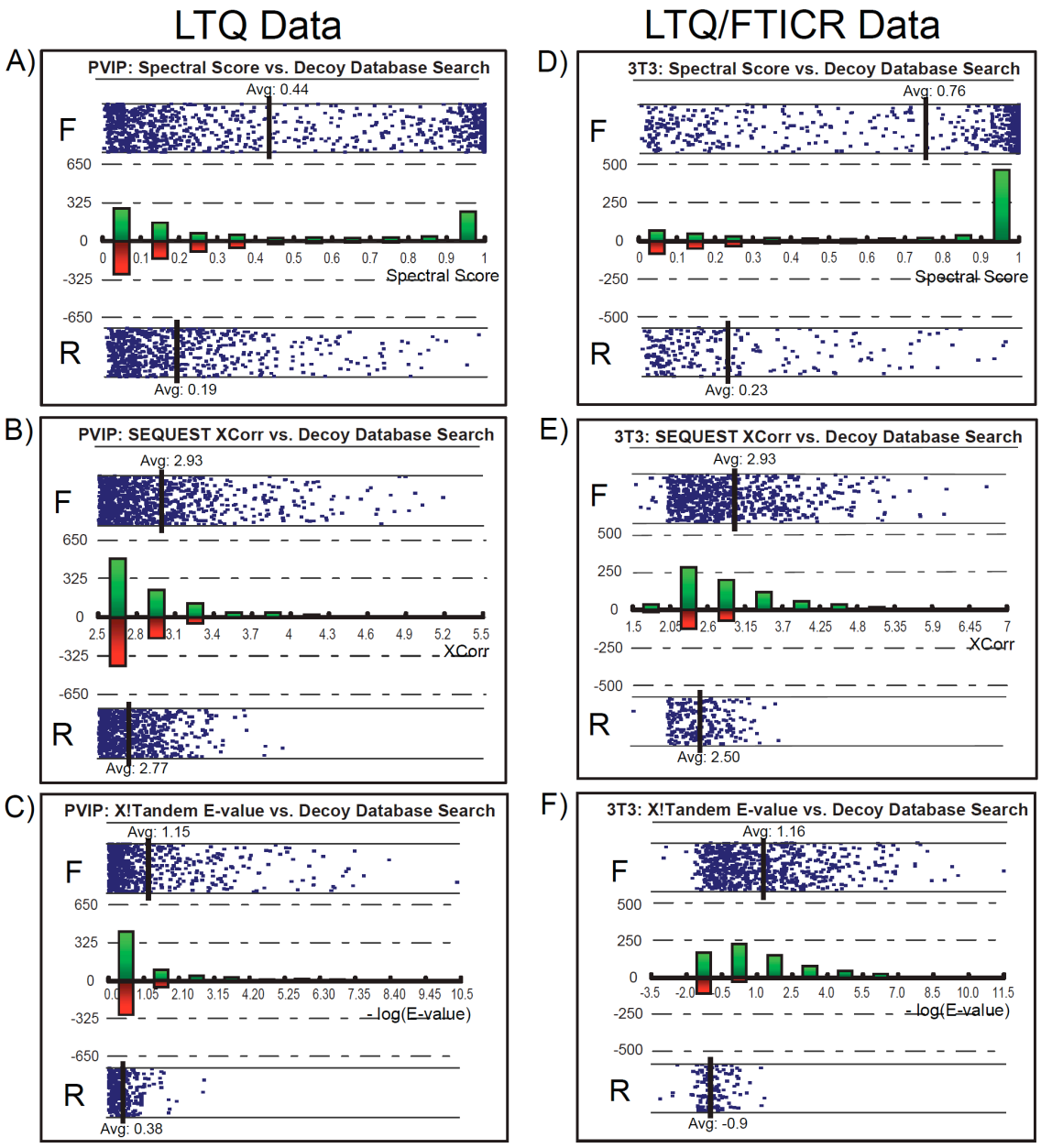
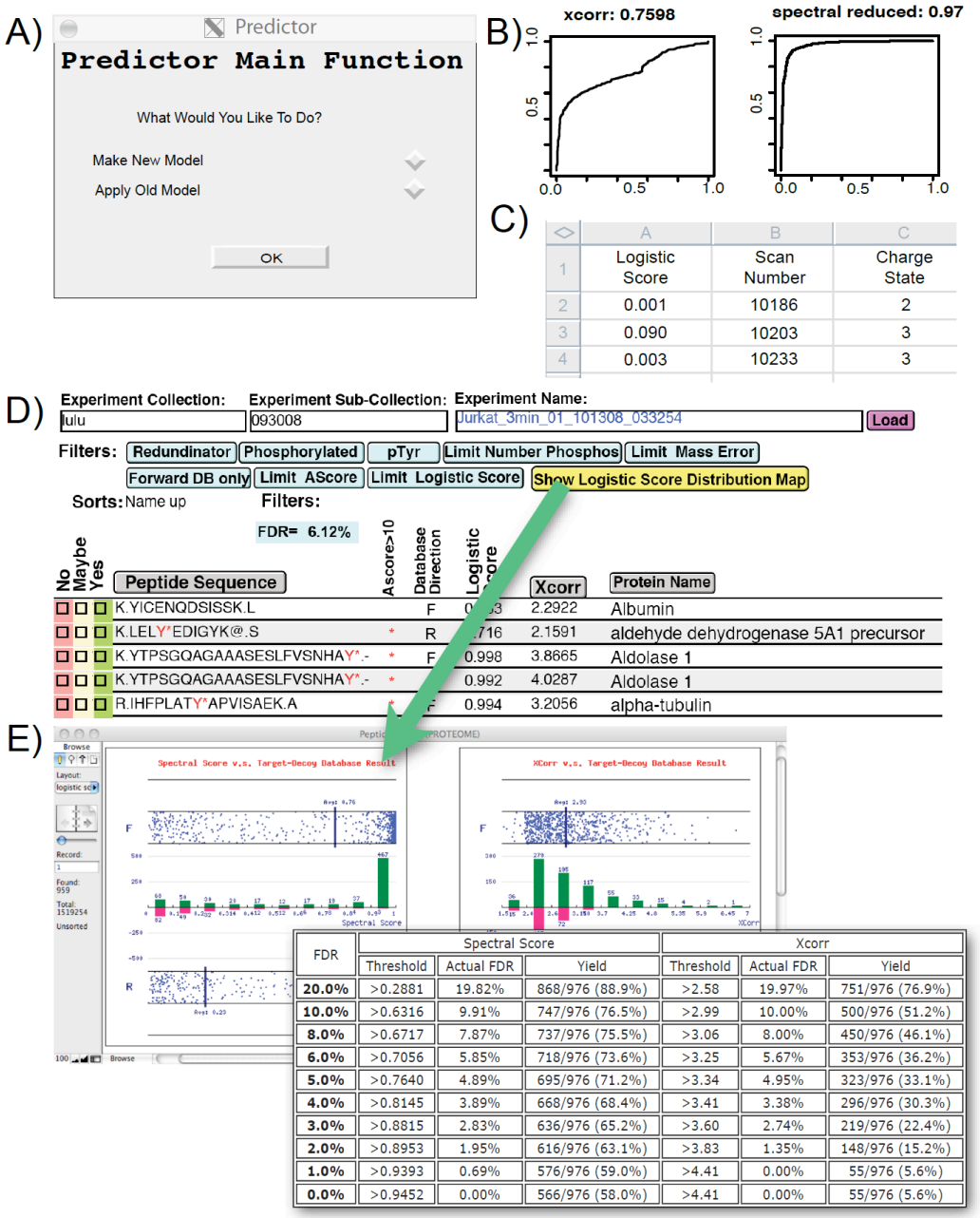


Figure 4.2: Performance of Spectral Model and XCorr evaluated with Decoy Database Approach. Scores are plotted against the protein database types, Forward (labeled as "F" in figures) and Reversed (labeled as "R" in figures) to demonstrate the distribution. Each data point in the figure represents an assigned peptide. A histogram of peptide counts was overlaid on the same figure with green bars representing forward hits and red bars representing reversed database hits. Spectral score is computed using the Spectral model trained on MCP5. Forward and reverse distribution for the PVIP dataset acquired on an LTQ mass spectrometer versus A) Spectral score, B) XCorr C) E-value calculated by X!Tandem. Forward and reverse distribution for the 3T3 dataset acquired on an LTQ/FTICR mass spectrometer with a 20 ppm mass error cutoff versus D) Spectral score, E) XCorr F) E-value calculated by X!Tandem.

To assist proteomics researchers with high throughput statistical analysis and generation of new statistical models, we have integrated model training and application of user-driven models within our automated data pipeline (Figure 4.3). In the analyses presented in this paper, training of our new logistic model was based upon the four sets of validated data mentioned above. A user may input any validated dataset from any type of mass spectrometer using any validation metrics to recompute model variables, tailoring the prediction to the user's needs and expectations through a flexible user interface. To facilitate new model building, a freely-downloadable, open-source software in the R programming language was developed to create new logistic models based either on user supplied validated training sets or datasets described in this manuscript (Figure 4.3A). When creating new models, this software also calculates ROC curves and the corresponding AUC for all models (Figure 4.3B). Any user-created logistic model may be applied to newly collected user data manually resulting in the output of logistic scores into a flat file for every peptide (Figure 4.3C). These user-generated models can also be integrated within our automated proteomic pipeline (HTAPP) that performs statistical analysis without user intervention after a dataset is newly acquired. In the automated mode, the newly calculated logistic score is deposited and viewable within our proteomic relational database PeptideDepot (Figure 4.3D). Within the database, a user may apply thresholds to the data, calculate FDR by the decoy database approach, and view plots of XCorr and spectral score versus decoy database direction for any filtered subset of experimental data (Figure 4.3D-E). These plots assist in the selection of appropriate logistic score thresholds for any user preferred FDR.



FDR	Spectral Score			Xcorr		
	Threshold	Actual FDR	Yield	Threshold	Actual FDR	Yield
20.0%	>0.2881	19.82%	868/976 (88.9%)	>2.58	19.97%	751/976 (76.9%)
10.0%	>0.6316	9.91%	747/976 (76.5%)	>2.99	10.00%	500/976 (51.2%)
8.0%	>0.6717	7.87%	737/976 (75.5%)	>3.06	8.00%	450/976 (46.1%)
6.0%	>0.7056	5.85%	718/976 (73.6%)	>3.25	5.67%	353/976 (36.2%)
5.0%	>0.7640	4.89%	695/976 (71.2%)	>3.34	4.95%	323/976 (33.1%)
4.0%	>0.8145	3.89%	668/976 (68.4%)	>3.41	3.38%	296/976 (30.3%)
3.0%	>0.8815	2.83%	636/976 (65.2%)	>3.60	2.74%	219/976 (22.4%)
2.0%	>0.8953	1.95%	616/976 (63.1%)	>3.83	1.35%	148/976 (15.2%)
1.0%	>0.9393	0.69%	576/976 (59.0%)	>4.41	0.00%	55/976 (5.6%)
0.0%	>0.9452	0.00%	566/976 (58.0%)	>4.41	0.00%	55/976 (5.6%)

Figure 4.3: Open-source R software for training logistic models and its application. In the manual mode: A) User selects whether to train a new model or apply an existing model to a new dataset. B) The software automatically trains logistic models based on user provided validated datasets. User validation is provided as a boolean value using any user preferred validation metric, such as expert manual spectral validation, or simple contrived proteomic mixtures. The software calculates an ROC curve and AUC for each newly trained model. C) When applying a model to a new dataset, the software outputs a validation score for each spectra identified. In the automatic mode: D) An R program trained on MCP5 using the Spectral model is implemented in HTAPP to perform statistical validation automatically. Results are imported into a FileMaker database and the false-discovery-rate is calculated. User can specify any logistic score threshold to adjust the FDR. Clicking on the yellow button brings a live figure showing the logistic score or XCorr vs. Decoy database distribution (E).

4.3.2 SpecNote for Manual Validation and Annotation

When manual validation of any peptide within the PeptideDepot database is necessary, the spectral annotation tool SpecNote is available. Manual validation involves the verification of the assigned peptide sequence and validation of any post-translational modification positions within the sequence. For both tasks, sequence coverage and spectral coverage (assigned ion current) are important parameters for successful analysis. SpecNote can provide critical information to accelerate this process. The graphical interface of SpecNote is shown in Figure 4.4A. The sequence coverage of the matched peptide is indicated in the lower left corner of the display area by the peptide sequence with colored bars above or below their respective amino acids representing the matching of theoretical b and y ions to observed peaks. When the mouse pointer is hovered over amino acid letters within the peptide sequence, fragment peaks corresponding to the selected amino acid are highlighted in the spectrum, allowing the user to locate specific peaks quickly. SEQUEST assigned phosphorylation site positions also need to be manually validated. SpecNote enables the user, by clicking on the modified amino acid and using the arrow key on the keyboard, to make a quick comparison of different repositioned post-translational modifications in both peak assignment and sequence coverage. A preference panel (Figure 4.4B), hidden in the normal view, can be displayed by pressing the space bar. By default, only b/y ions and user annotations are labeled in the spectrum. Using this preference panel, other ion types, such as c, z, a, and x, as well as neutral loss of water, ammonia, and phosphate can also be labeled. Also a user may adjust the labeling threshold, the divisions of x-axis, and alter the sequence or modification site of the peptide.

SpecNote also incorporates novel features not present in other similar tools such as a snap-to-peak function. A normal MS/MS spectrum contains many peak assignments, making it impossible to display all detailed information at the same time. With the snap-to-peak feature, the program senses the current mouse position and, if it falls within a predefined distance from an MS/MS fragment peak, the mouse pointer snaps to that peak. A window then pops up displaying details about that peak, such as m/z, relative abundance, and suggested theoretical ion assignment with the associated mass error. This feature removes the need for the proteomic end-user to zoom in and out to retrieve that same information, maintaining user orientation.

SpecNote accelerates the manual validation process by performing numerous calculations for the user and integrating calculated results within the assigned spectra. Traditionally, the user would manually calculate all possibilities to match unidentified fragments. Such procedures reduce manual validation throughput. SpecNote takes less than a second to assign all peaks to theoretical fragments by using an automatic peak assignment function to calculate mass differences (Δmass) between the observed masses and any potential theoretical fragment masses, including b/y/c/z/a/x ions. Since neutral losses of precursor or fragment ions are widely observed in CID spectra, the algorithm also searches for neutral losses of water, ammonia, and phosphate (in the case of phosphorylated peptides) from all applicable fragments and the precursor ion. All possible charge states are considered. After calculation of all possible assignments, the fragment is automatically assigned using the following hierarchy: 1) b and y ions are preferred for CID spectra by default, 2) Δmass is minimized, and 3) the number of neutral losses is minimized. Clicking the peak allows the user to compare amongst all calculated

theoretical assignments for a given peak including mass errors (Figure 4.4C) and select one of them or even enter a manual annotation if the user disagrees with computer suggested assignments, increasing accuracy and efficiency. On average, SpecNote can save at least five minutes per spectrum compared to printing the spectra and labeling the peaks manually.

SpecNote allows the user to export PDF reports for selected spectra along with all the assignments and annotations by incorporating iText (http://www.lowagie.com/iText/), a free Java-PDF library. iText generates PDF in vector mode, so the file size is only around 8 kilobytes per spectrum which is convenient for distributing data between labs.

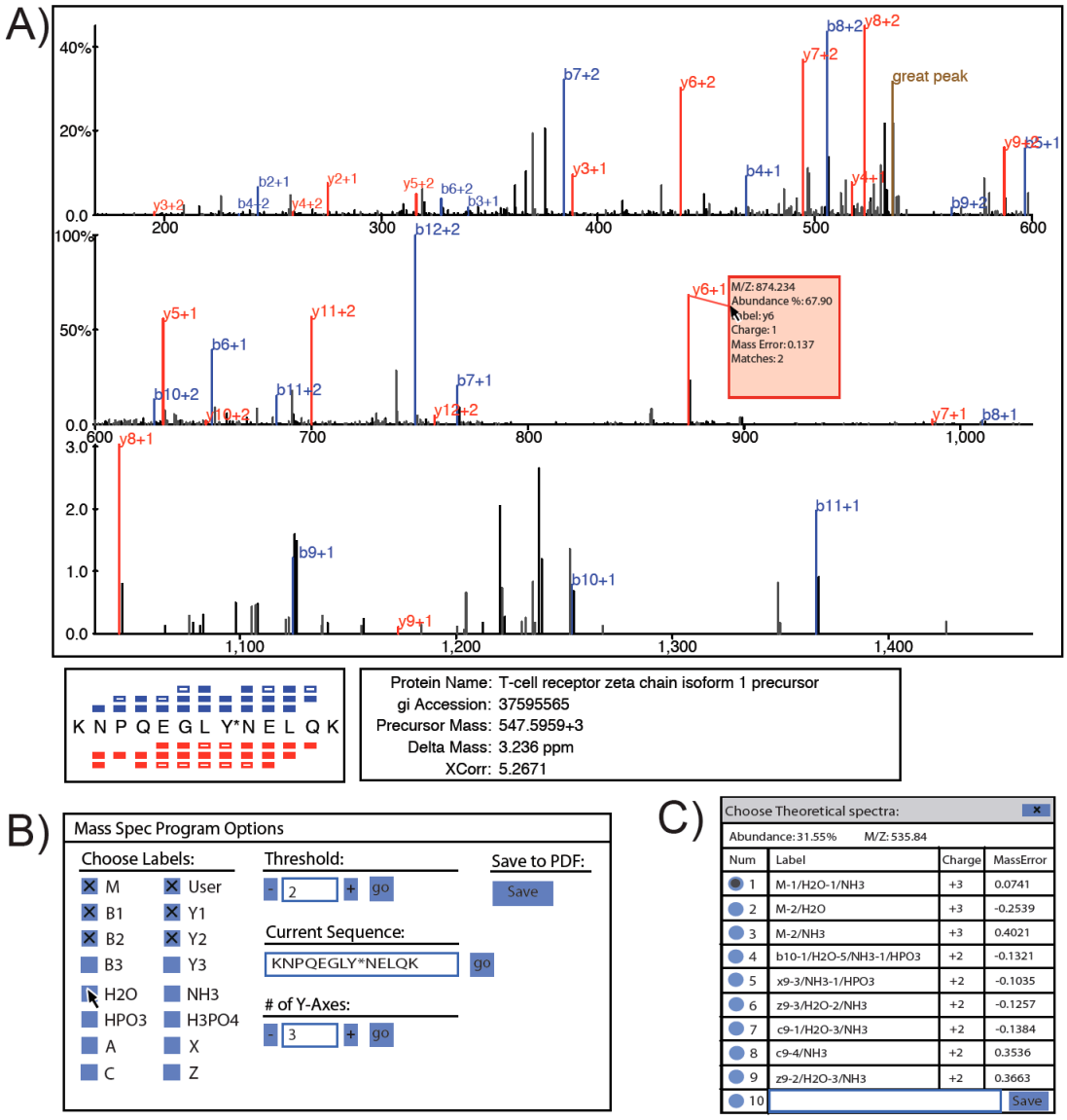


Figure 4.4: User interface of SpecNote. A) Main display area showing essential peptide metadata and a spectrum with peaks auto-assigned to peptide fragments. Sequence coverage is illustrated with b-series ions on top and y-series ions at bottom. Detailed information such as protein name, XCorr, delta mass is provided and linked to NCBI website. B) Preference panel allows a user to determine which ions to show by selecting ion types and intensity threshold. Assigned sequence can be changed by resubmitting a modified sequence. C) For each peak in the spectrum, a user can reassign it to other candidate fragments or input a customized annotation. All user annotations are stored automatically within MySQL proteomic database.

4.4 CONCLUDING REMARKS

Recent innovations in multi-dimensional LC/MS proteomic methods have led to a deluge of data generation in proteomics. The ability to efficiently discern the true assignment of MS/MS spectra to peptide sequence is essential in this context.

Proteomic researchers sharply differ in the appropriateness of certain methods of data validation. Many investigators perform simple thresholding on SEQUEST parameters while approximating false discovery rates with decoy database search [37]. Other researchers prefer the development of statistical models based on simple protein mixture digests with the assumption that true positive hits only result from the known proteins in the mixture with hits to any other protein defined as false positives. The popular PeptideProphet algorithm was trained with this type of approach [12]. Other researchers, weary of the possibility of unexpected contaminating proteins present in these contrived mixtures, prefer a manual interpretation of MS/MS spectra. Our logistic spectral score is adaptable to any user preferred validation philosophy with model training implemented as a fully supported software feature.

An illustration of the power of model training in creating optimal models for certain proteomic workflows is the analysis of phosphoprotoemic datasets. Although the development of entirely new statistical approaches can optimize the yield of phosphopeptides at a user selected FDR [11], we show here that our logistic spectral model trained with validated phosphoproteomic datasets (MCP5 and PVIP) outperforms logistic spectral models trained with the BSA unphosphorylated dataset. For example, the AUC for cross application of PVIP and MCP5 trained models was 0.920 and 0.897

compared to 0.908 and 0.837 for application of the unphosphorylated BSA trained model onto PVIP and MCP5.

One criticism of user-driven model training is the difficulty of comparison of the predictions from multiple user created models from different labs with each other. To address this criticism, we have integrated the standardized estimation of false discovery rate by the decoy database approach as a central component of our fully automated data analysis pipeline. By providing a graphical representation of the distribution of spectral scores relative to forward and reversed database hits, a user may compare model performance from different user-created models and select a user preferred FDR. Along with spectral score thresholds, a user may supply these estimated FDR from decoy database approach to provide a universal, unbiased assessment of the quality of proteomic data submitted for publication to scientific journals. Furthermore, the statistical models may be easily distributed as supplemental material when data is submitted for publication in the form of a single binary R data file.

Overall, the combination of user driven logistic spectral models, full automation of statistical analysis within high-throughput proteomic workflows and accelerated manual spectral annotation within the PeptideDepot proteomic relational database increases both the efficiency of proteomic workflows along with increased yield of confident peptide assignments.

Currently, the software described here is designed around a SEQUEST workflow. To adapt the software to a new search engine such as X!Tandem or Mascot, the R software and the SpecNote annotation software would require updated parsing of the database search results. Additionally the logistic spectral model would need to be trained on the

output variables of the new search algorithm. In the future, logistic spectral models may also be trained with variables from multiple search algorithms to collate database search scores such as XCorr, E-value, and MOWSE score into a unified logistic spectral score. These small modifications could easily be implemented within the source code of our software.

4.5 REFERENCE

1. Eng, J.K., A.L. McCormack, and J.R. Yates, *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.* J Am Soc Mass Spectrom, 1994(5): p. 976-989.

2. Craig, R. and R.C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2310-6.

3. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.

4. Clauser, K.R., P. Baker, and A.L. Burlingame, *Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching.* Anal Chem, 1999. **71**(14): p. 2871-82.

5. Nesvizhskii, A.I. and R. Aebersold, *Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS.* Drug Discov Today, 2004. **9**(4): p. 173-81.

6. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

7. Taylor, J.A. and R.S. Johnson, *Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry.* Anal Chem, 2001. **73**(11): p. 2594-604.

8. Ma, B., et al., *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2337-42.

9. Tabb, D.L., A. Saraf, and J.R. Yates, 3rd, *GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model.* Anal Chem, 2003. **75**(23): p. 6415-21.

10. Tanner, S., et al., *InsPecT: identification of posttranslationally modified peptides from tandem mass spectra.* Anal Chem, 2005. **77**(14): p. 4626-39.

11. Lu, B., et al., *Automatic validation of phosphopeptide identifications from tandem mass spectra.* Anal Chem, 2007. **79**(4): p. 1301-10.

12. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.

13. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nat Biotechnol, 2006. **24**(10): p. 1285-92.

14. Sun, S., et al., *Improved validation of peptide MS/MS assignments using spectral intensity prediction.* Mol Cell Proteomics, 2007. **6**(1): p. 1-17.

15. Anderson, D.C., et al., *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, 2003. **2**(2): p. 137-46.

16. Higdon, R., et al., *LIP index for peptide classification using MS/MS and SEQUEST search via logistic regression.* OMICS, 2004. **8**(4): p. 357-69.

17. Kall, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nat Methods, 2007. **4**(11): p. 923-5.

18. Razumovskaya, J., et al., *A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST.* Proteomics, 2004. **4**(4): p. 961-9.

19. Sun, W., et al., *AMASS: software for automatically validating the quality of MS/MS spectrum from SEQUEST results.* Mol Cell Proteomics, 2004. **3**(12): p. 1194-9.

20. Gentzel, M., et al., *Preprocessing of tandem mass spectrometric data to support automatic protein identification.* Proteomics, 2003. **3**(8): p. 1597-610.

21. Moore, R.E., M.K. Young, and T.D. Lee, *Method for screening peptide fragment ion mass spectra prior to database searching.* J Am Soc Mass Spectrom, 2000. **11**(5): p. 422-6.

22. Sadygov, R.G., H. Liu, and J.R. Yates, *Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases.* Anal Chem, 2004. **76**(6): p. 1664-71.

23. Wu, F.X., et al., *RT-PSM, a real-time program for peptide-spectrum matching with statistical significance.* Rapid Commun Mass Spectrom, 2006. **20**(8): p. 1199-208.

24. Payne, S.H., et al., *Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis.* J Proteome Res, 2008. **7**(8): p. 3373-81.

25. Olsen, J.V., et al., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.* Cell, 2006. **127**(3): p. 635-48.

26. Cao, L., et al., *Quantitative time-resolved phosphoproteomic analysis of mast cell signaling.* J Immunol, 2007. **179**(9): p. 5864-76.

27. Hoffert, J.D., et al., *Quantitative phosphoproteomics of vasopressin-sensitive renal cells: regulation of aquaporin-2 phosphorylation at two sites.* Proc Natl Acad Sci U S A, 2006. **103**(18): p. 7159-64.

28. Lehtinen, M.K., et al., *A conserved MST-FOXO signaling pathway mediates oxidative-stress responses and extends life span.* Cell, 2006. **125**(5): p. 987-1001.

29. Eddes, J.S., et al., *CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies.* Proteomics, 2002. **2**(9): p. 1097-103.

30. Klimek, J., et al., *The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools.* J Proteome Res, 2008. **7**(1): p. 96-103.

31. Ficarro, S.B., et al., *Automated immobilized metal affinity chromatography/nano-liquid chromatography/electrospray ionization mass spectrometry platform for profiling protein phosphorylation sites.* Rapid Commun Mass Spectrom, 2005. **19**(1): p. 57-71.

32. Brill, L.M., et al., *Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry.* Anal Chem, 2004. **76**(10): p. 2763-72.

33. Secrist, J.P., et al., *Stimulatory effects of the protein tyrosine phosphatase inhibitor, pervanadate, on T-cell activation events.* J Biol Chem, 1993. **268**(8): p. 5886-93.

34. Tanaka, S., T. Ito, and J.R. Wands, *Neoplastic transformation induced by insulin receptor substrate-1 overexpression requires an interaction with both Grb2 and Syp signaling molecules.* J Biol Chem, 1996. **271**(24): p. 14610-6.

35. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

36. Link, A.J., et al., *Direct analysis of protein complexes using mass spectrometry.* Nat Biotechnol, 1999. **17**(7): p. 676-82.

37. Villen, J., et al., *Large-scale phosphorylation analysis of mouse liver.* Proc Natl Acad Sci U S A, 2007. **104**(5): p. 1488-93.

Chapter 5

HIGH-THROUGHPUT, QUANTITATIVE ANALYSIS OF INSULIN SIGNALING PATHWAY

5.1 INTRODUCTION

Cellular behaviors are extensively regulated by various signaling pathways, among which insulin signaling plays a critical role in nutrition metabolism and cell proliferation/apoptosis [1]. The canonical insulin signaling pathway diagram is illustrated in Figure 5.1. Previous studies suggests that the insulin signaling pathway is initiated via the binding of insulin protein to the extracellular domain of insulin receptor (IR) α subunit [2]. Insulin receptor β subunit is then immediately autophosphorylated, resulting in elevated enzymatic activity toward its substrate, insulin receptor substrate (IRS) family of proteins. Activated IRS proteins serves as a messenger to mediate and transmit insulin receptor signaling to downstream networks, by recruiting several intracellular signaling molecules containing Src homology 2 (SH2) binding motifs [3, 4]. Several important downstream signaling cascades are known to be associated with this process. IRS family activates PI3K/Akt cascade by phosphorylation of p85 regulatory subunit of phosphatidylinositol 3-kinase (PI3K) [5, 6], leading to various critical physiological effects. Akt is a serine/threonine kinase that inhibits lipolysis by activating phosphodiesterase 3B (PDE3B) [7]; promotes glycogen synthesis and fatty acid synthesis levels by inhibiting glycogen synthase kinase 3 (GSK-3) [8]; protects cell against apoptosis by inhibiting Bad [9]. Translocation of Akt to the nucleus inhibits Forkhead box O1 (FOXO1), Forkhead box O3A (FOXO3A) and Forkhead box O4 (FOXO4), and turns off programmed cell death [10, 11]. On another branch of the insulin signaling pathway, IRS family proteins regulate the Ras/mitogen-activated protein kinase (MAPK) cascade by recruiting Grb2 and/or SHP2 [12-18]. This activation finally initiates the mitogenic response [19].

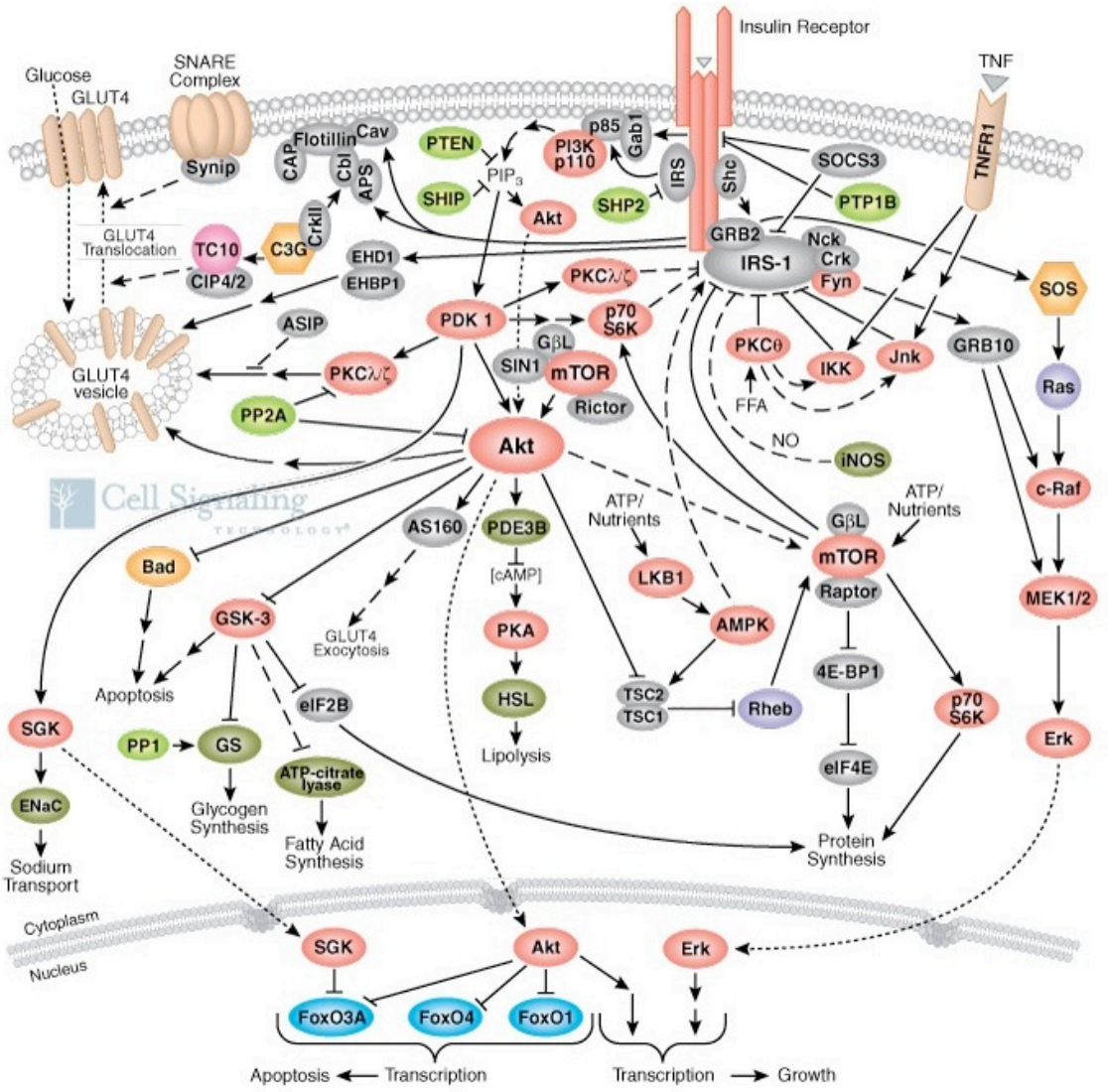


Figure 5.1: Canonical diagram of the insulin signaling pathway. (Figure from Cell Signaling Technology)

It is a fair assumption that many malicious tumors are associated with a dysfunctional insulin signaling pathway, including hepatocellular carcinoma (HCC). HCC is a primary type of liver cancer that causes severe death within the human population. In most HCC tumor tissues and cell lines, overexpression of IRS-1 protein has been identified [20]. Based on the canonical insulin signaling pathway, elevated IRS-1 level could result in over-sensitive response to external stimulus, which may lead to un-regulated cell proliferation. In order to study how IRS-1 mediates extracellular signals and interacts with other proteins, several

NIH3T3 cell lines have been established by transfection of mutant human IRS-1 (hIRS-1) transcripts [18]. (Figure 5.2) These cell lines have been characterized previously and show a neoplastic transformation phenotype [18, 21]. Therefore, this model system to study the insulin signaling pathway in HCC can reveal important details of the molecular basis of this disease.

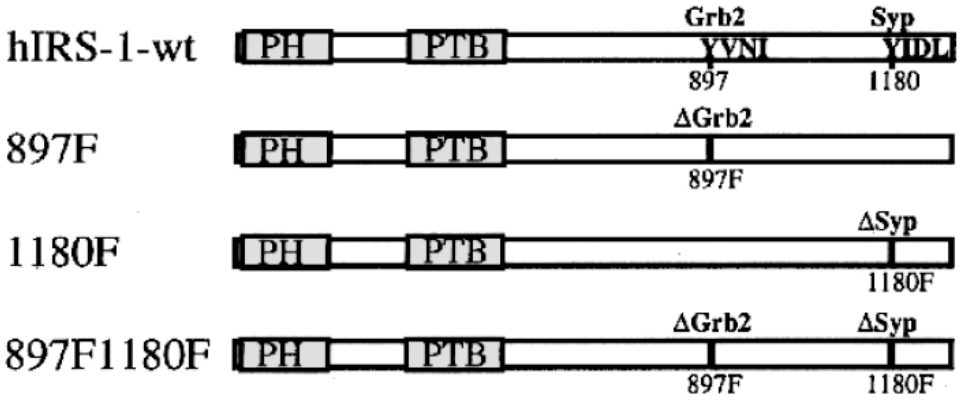


Figure 5.2: Schemetic representation of wild-type and mutated hIRS-1 proteins overexpressed in NIH3T3 cells. (Figure from Tanaka et al [18])

In this chapter, we investigate the global phosphorylation in NIH3T3 cells expressing Y1180F mutant IRS-1 through a time course of insulin stimulation. Data is analyzed using the previously described high-throughput proteomic pipeline (Chapter 2-4) to unravel quantitative perturbations of the phosphoproteome in this mutant cell through a time course of insulin receptor stimulation.

5.2 MATERIALS AND METHODS

5.2.1 Cell Culture and Stimulation

Wild type NIH3T3 and IRS-1 Y1180F mutant NIH3T3 cells were described previously. [18] We obtained these cells from the lab of Prof. Wands at Brown University. Two cell lines, NIH3T3 cells with overexpression of human IRS1 protein (NIH3T3-hIRS1) and NIH3T3 cells with overexpression of human IRS1 protein with SHP2 binding site mutated (NIH3T3-hIRS1 Y1180F) were maintained in Dulbecco's modified Eagle's medium (DMEM) (Sigma) containing either $^{12}C_6$, $^{14}N_4$ Arg and $^{12}C_6$, $^{14}N_2$ Lys (Sigma) or $^{13}C_6$, $^{15}N_4$ Arg and $^{13}C_6$, $^{15}N_2$ Lys (Cambridge Isotope Laboratories, Andover MA) supplemented with 10% heat inactivated dialyzed fetal bovine serum (Sigma), 2mM L-glutamine (Sigma), 400 μg/mL G418 (Sigma), 100 μM non-essential amino acids (Invitrogen, Carlsbad CA) in a humidified incubator with 5% CO_2 at 37 °C for 7 cell doublings. The concentration of Lys and Arg used in SILAC labeling of NIH3T3 cells in experiments described here was 37 mg/L and 23 mg/L, respectively. Cells were starved in DMEM with 0.1% bovine serum albumin for 24 hours after reaching 95% confluence.

Insulin stimulation of NIH3T3 cells was performed as described previously [22]. Briefly, starved cells were washed once with 4 °C phosphate buffer saline (PBS), and reconstituted at a concentration of 15 ml PBS per 15cm cell culture dish. For each time point, $1x10^7$ cells were treated with 100 nM insulin (Sigma) and incubated at 37 °C for 0, 0.5, 1, 3, 5, 10, or 30 minutes.

5.2.2 Protein Harvest and Digestion

To stop insulin stimulation, cells were treated similarly as described elsewhere [23]. Cells were placed in lysis buffer (8 M urea, 1 mM sodium orthovanadate, and 100 mM ammonium bicarbonate, pH 8.0) and incubated for 20 min at 4 °C. Lysates were then cleared at 12,000 g

for 15 min at 4 °C, and protein concentrations were measured by the DC Protein Assay (Bio-Rad, Hercules, CA). Cell lysates from NIH3T3-hIRS1 and NIH3T3-hIRS1 Y1180F were combined at a 1:1 protein concentration ratio and reduced with 10 mM DTT for 1 hr at 56 °C, followed by alkylation with 55 mM iodoacetamide for 1 hr at room temperature in dark. Proteins were then diluted 5 fold with 100 mM ammonium bicarbonate (pH 8.9) and digested with sequencing grade modified trypsin (Promega, Madison, WI) at a 1:100 (w/w) trypsin:protein ratio overnight at room temperature. Tryptic peptides were acidified to pH 2, cleared at 2000 g for 10 min at 22 °C, desalted using C18 Sep-Pak plus cartridges (Waters, Milford, MA), and lyophilized in a Speed Vac plus (Thermo Fisher Scientific, Waltham, MA).

5.2.3 SCX Fractionation and TiO_2 Enrichment

Offline strong cation exchange (SCX) was performed on an Agilent 1200 HPLC system to fractionate peptide samples. Each timepoint was dissolved in water with 0.1 M acetic acid and loaded onto a custom packed SCX column (0.75mm ID x 25cm) of Poly Sulfoethyl A (300 Å), 5μm resin (Nest Group, Southborough, MA). Peptides were eluted from the column using a 12-step salt gradient with buffer A (30% acetonitrile, 5mM NaH_2PO_4, pH 3.0) and buffer B (30% acetonitrile, 5mM NaH_2PO_4, pH 3.0, 500mM $NH_4CH_3COO^-$). Salt step used for fractions collected was: fractions 1-10, step-wise 0-20% buffer B, fraction 11, 50% buffer B and fraction 12, 100% buffer B. Each fraction was then subjected to TiO_2 phosphopeptide enrichment as described previously with a few modifications [24]. Eluents were diluted 5:1 with 30 g/L 2,5-dihydroxybenzoic acid (DHB) in 80% MeCN and 0.1% trifluoroacetic acid (TFA) and incubated with 10 mg of titanium beads (GL Science, Torrance, CA) pre-washed

with 5 mg/mL DHB in 80% MeCN for 1 hour. Beads were then washed twice with 200 µL of 50% MeCN in water with 0.2% TFA. Phosphopeptides were eluted using 100 µ L of ammonium solution in 40% MeCN, pH 10.5 for 15 minutes.

5.2.4 Mass Spectrometric Analysis

Enriched phosphopeptides were injected into the mass spectrometer (LTQ-FT; Thermo Fisher Scientific) through an analytical column (360 µm OD X 75 µm ID fused silica with 12 cm of 5 µm Monitor C18 particles with an integrated ~5 µm ESI emitter tip fritted with 3 µm silica; Bangs Laboratories) with a reversed-phase gradient (0-70% MeCN with 0.1 M acetic acid in 30 min). Static peak parking was performed via flow rate reduction from 200 nl/min to ~40 nl/min when peptides began to elute as judged from a BSA peptide scouting run, as described previously [25]. Using a split flow configuration, an electrospray voltage of 2.0 kV was applied as described [26]. Spectra were collected in positive ion mode and in cycles of one full MS scan in the FT (m/z: 400-1800), followed by data-dependent MS/MS scans in the LTQ (~0.3 s each) sequentially of the five most abundant ions in each MS scan with charge state screening for +1, +2, +3 ions and dynamic exclusion time of 30 s. The automatic gain control was 1,000,000 for the FTMS scan and 10,000 for the ion trap MS (ITMS) scans. The maximum ion time was 100 ms for the ITMS scan and 500 ms for the FTMS full scan. FTMS resolution was set at 100,000.

MS/MS spectra were searched against the mouse National Center for Biotechnology Information non-redundant protein database using the Mascot algorithm (ver 2.2.1) [27]. Peak lists were generated using extract_msn.exe provided by ThermoFisher using a mass range of 600-4500, precursor ion tolerance (for grouping) of 0.005 AMU, minimum ion

count of 5, group scan of 0, minimum group count of 1. The NCBI mouse database was appended with its reversed version (decoy database) [28]. Mascot search was performed with the following parameters: trypsin enzyme specificity, 2 possible missed cleavages, 0.2 Da mass tolerance for precursor ions, 0.5 Da mass tolerance for fragment ions. Search parameters specified a differential modification of phosphorylation (+79.9663 Da) on serine, threonine, and tyrosine residues and a static modification of carbamidomethylation (+57.0215 Da) on cysteine. Search parameters also included a differential modification for arginine (+10.00827 Da) and lysine (+8.01420 Da) amino acids. To provide high confidence phosphopeptide sequence assignments, Mascot results were filtered with ion score (>25), precursor mass error (<20 ppm), a logistic spectral score that assessed MS/MS spectral quality (> 0.5015), non-redundant phosphopeptides, and proteins with descriptors of "unnamed" or "unknown" removed, to reach 1% false discovery rate estimated by the decoy database approach [28]. To validate the position of the phosphorylation site, the Ascore algorithm [29] was applied to all data, and the reported phosphorylation site position reflects the top Ascore prediction. All data were processed by HTAPP platform described previously in an automated fashion.

5.3 RESULTS

5.3.1 SCX/TiO_2 Approach Revealed 2201 Phosphorylation Sites in NIH3T3 Cells

All peptide hits identified in a total of 91 LC/MS runs (7 timepoints times 13 fractions per timepoint) were assembled into a quantitative comparison timecourse. At 1% FDR rate estimated by decoy database, we discovered a total of 2201 phosphorylation sites on 1194

proteins, among which 1862 (84.6%) were on serine, 299 (13.6%) were on threonine and 40 (1.8%) were on tyrosine (Figure 5.3 A). These numbers were consistent with previous estimated phosphorylation site distributions [30]. We found that the TiO_2 enrichment method produced predominantly singly phosphorylated peptides (Figure 5.3 B).

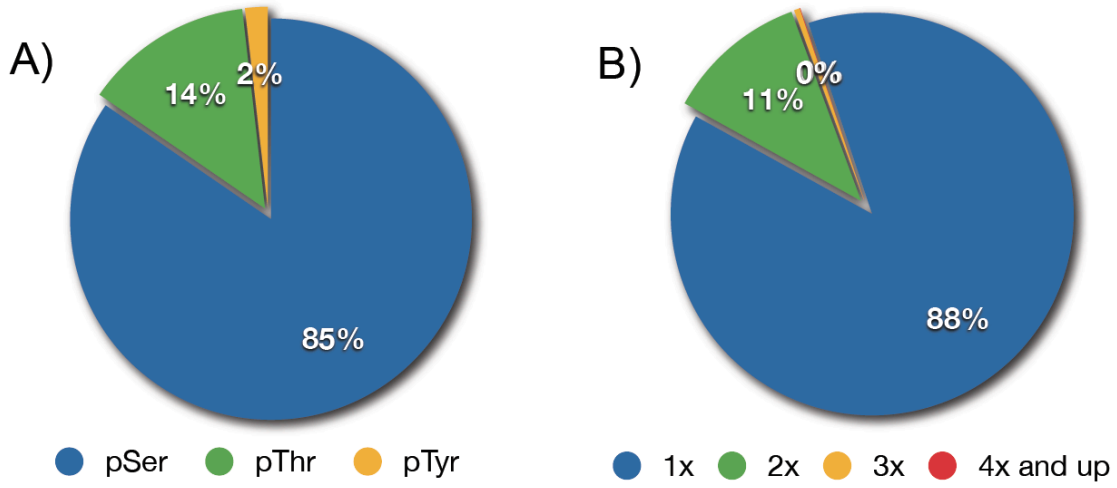


Figure 5.3: Results of large-scale phosphoproteomic analysis of insulin-stimulated NIH3T3 cells. A) Residue distribution of phosphopeptides; B) Distribution of singly, doubly, and multiply phosphorylated peptides.

5.3.2 Quantitative Time-Resolved Proteomic Data Revealed Different Regulation Patterns in NIH3T3-hIRS1 and NIH3T3-hIRS1 Y1180F

Using a hybrid approach combining both a label-free and SILAC quantitation strategy (Figure 1.5), identified phosphopeptides were quantified and data assembled in the experiment comparison view within PeptideDepot. We detected many known and novel phosphorylation sites on IGF-II receptor, IRS1 and IRS2, indicating that insulin signaling pathway was triggered after treatment with insulin. Data show the protein activation after 0 minutes. (Figure 5.4)

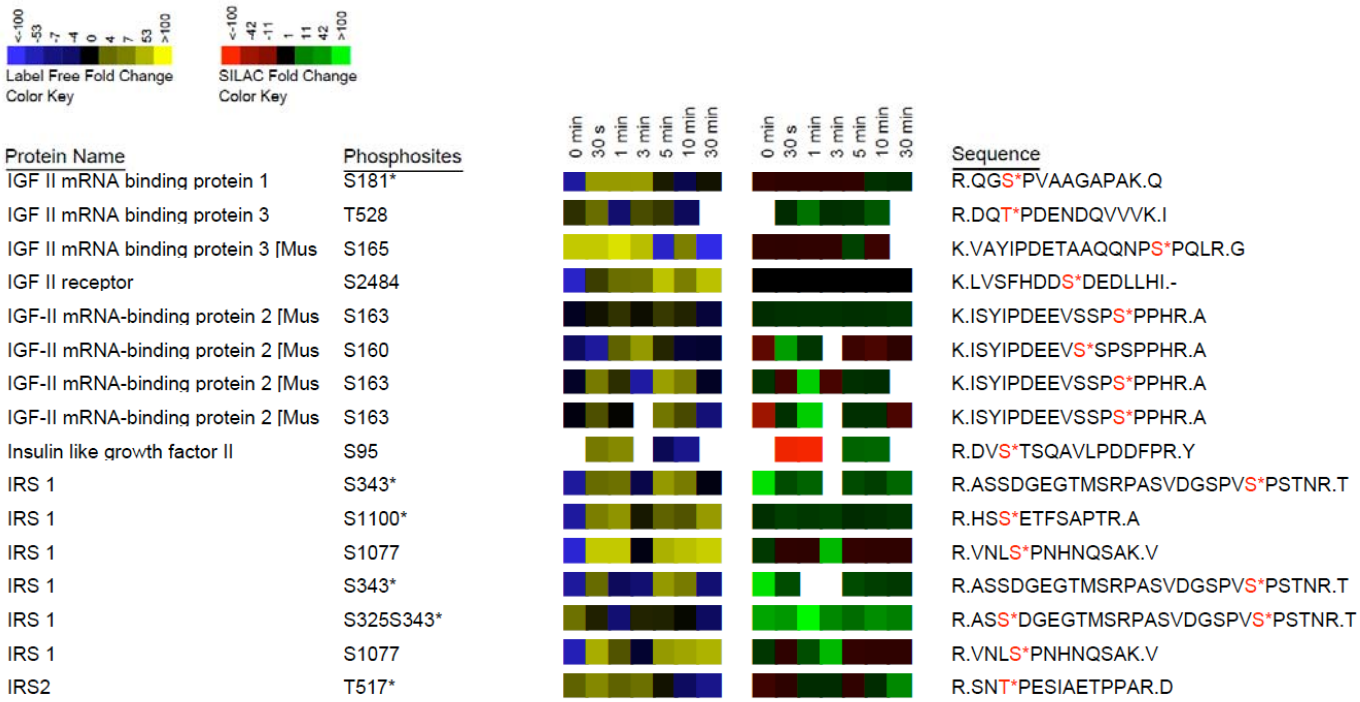
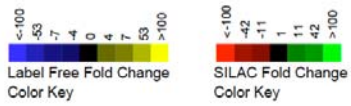


Figure 5.4: Activation of insulin/IGF related proteins in insulin-stimulated NIH3T3 cells. The heatmap on the left shows the temporal profile of phosphorylations in NIH3T3-hIRS1 cells. The heatmap on the right shows the ratio of phosphorylation abundance in NIH3T3-hIRS1 Y1180F versus NIH3T3-hIRS1. Color legends are indicated on the top-left.

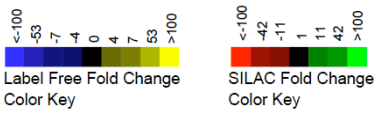
SHP2 (also known as Syp, SHPTP2, PTP1D, SHPTP3, or PTP2C) is a protein tyrosine phosphotase that contains two adjacent SH2 domains in addition to a catalytic phosphatase domain. SHP2 binds to insulin receptor substrate 1 to become activated [15, 31]. In the NIH3T3-hIRS1 Y1180F cell line, a point mutation was created to replace Y with F, leading to failure of IRS1 to activate SHP2. Previous studies suggest that SHP2 is a positive downstream regulator of insulin signaling pathways [32] and removal of SHP2 binding site leads to decreased MAPK activity in response to insulin stimulation comparing to NIH3T3-hIRS [18]. We summarized our results into three categories: phosphorylation sites that were down-regulated in NIH3T3-hIRS1 Y1180F comparing to NIH3T3-hIRS1 (Figure 5.5); phosphorylation sites that were up-regulated in NIH3T3-hIRS1 Y1180F comparing to

NIH3T3-hIRS1 (Figure 5.6); and phosphorylations that didn't show obvious changes between the two cell lines (Figure 5.7).



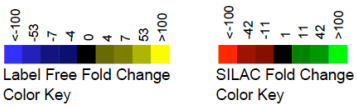
Protein Name	Phosphosites	Sequence
100 kDa thyroid hormone receptor	T842	R.LLSSSDDDANILSSPT*DR.S
1700023B02Rik protein [Mus	S202	R.NLTANDPSQDYVAS*DCEEDPEVEFLK.S
6330414O09Rik protein [Mus	S89	R.KAS*PEPEGETAGK.M
AA589382 protein [Mus musculus]	S921S934	R.S*NSFSDEREFSAPS*TPTGTLEFAGGDAK.G
ACINUS	S479	R.SLS*PLSGTTDTK.A
ASAP1	S406	K.QAS*EDSNDLTPTLPETPVPLPR.K
Ataxin 2 domain protein	S291*	K.LQPSSS*PETGLDPFPSR.I
BAT2 domain containing 1	T290	R.STT*PTSSPFR.A
BCL2 associated athanogene 3	S380S390	K.VSSAPIPCPS*PSPAPSAVPS*PPKNVAAEQK.A
Cadherin 11	S788	K.LADLYGS*KDTFDDDS.-
Caldesmon 1	S249	K.LKQTENAFS*PSR.S
Catenin, delta 1	S268*	R.VGGS*SVDLHR.F
CDC2	Y15*	K.IGEGTY*GVVYK.A
Collapsin response mediator protein	T509*	R.GLYDGPVCEVSVT*PK.T
Cyclin dependent kinase inhibitor 1A	S15	K.VYLS*PGSR.S
Dynamin 1	S774S778*	R.RS*PTSS*PTPQR.R
Eplin	S212	R.TSSLPESS*PSK.T
expressed sequence AA415817	T687T691	K.VQGTGVT*PPPT*PLGTR.S
expressed sequence AI131757 [Mus	S46	R.SHS*LPNSLDYAQASER.G
expressed sequence C87860 [Mus	T470	R.DATIPVESDT*DDEGAPR.I
expressed sequence C87860 [Mus	S468T470	R.DATIPVES*DT*DDEGAPR.I
Family with sequence similarity 54,	S103	R.NAS*VPNLR.G
Filamin A, alpha	S968	K.SPFSVGVSPS*LDLSK.I
FMRP interacting protein, 82-kD	Y218	K.GADNDGSGSESGY*TTPK.K
Forkhead box protein O3A	S297	K.WPGSPTS*R.S
formin binding protein 1 [Mus	S296	R.TVS*DNSLSSSK.E
Grb4	S90*	R.DAS*PTPSTDAEYPANGSGADR.I
Heat shock protein 27	S82*	R.QLS*SGVSEIR.Q
intermediate filament protein nestin	S1522	K.SPAS*PKWEQAGEQR.L
Iroquois homeobox protein 1	S267	R.DQSS*PLSAAETLK.S
JNK-binding protein JNKBP1 [Mus	S1258	R.SIS*VGENPGLATEPQAPAPIR.I
Lamin A/C	S22*	R.SGAQASSTPLS*PTR.I
Lamin A/C	S17*	R.SGAQAS*STPLSPTR.I
Lim and SH3 protein 1	T104*	K.GFSVVADT*PELQR.I
Myosin 9B	T1346	R.ATGAALT*PTEER.R

(Continued from the previous page)



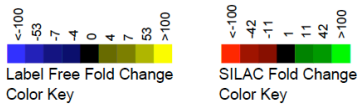
Protein Name	Phosphosites	Sequence
Myosin heavy chain 9, nonmuscle	S512*	K.GTGDCS*DEEVDGKADGADAKAAE.-
Myotubularin related protein 2	S8	K.SSS*CESLGAQLPAAR.L
N-myristoyl transferase 1	S47*	R.SGLS*PANDTGAK.K
Osmotic stress protein 94	T593	R.QLT*QDLLNSYIENEGK.M
p21/Cdc42/Rac1-activated kinase 1	T229*	R.DVATSPISPTENNTT*PPDALTR.N
PARD3	S1221	R.QYSS*LPR.Q
pericentriolar material-1 [Mus	S65S69	R.VTNAIS*PESS*PGVGR.R
phosphatidylinositol 4-kinase type II-	S47	R.VAAAAGSGPS*PPCSPGHDRER.Q
Plectin 1	S4613*	K.GYYS*PYSVSGSGSTAGSR.T
Poly(rC) binding protein 1	S190*	R.VMTIPYQPMPASS*PVICAGGQDR.C
Protein tyrosine kinase 9	T349	R.LIRGPAEAEATT*D.-
RAB3A interacting protein	T217	K.TLVLSSSPT*SPTQEPLAAAK.T
RAD	S301	K.SKS*CHDLSVL.-
Ral guanine nucleotide dissociation	S506S509	R.VIEPPAAS*CPS*SPR.I
Ribosomal protein S6 kinase alpha 3	S608	R.NQS*PVLEPVGR.S
serine/arginine repetitive matrix 2	S1209	K.NSGPVSEVNTGFS*PEVK.E
SH3 multiple domains 1	S405	K.LAQGS*PAVAR.I
Solute carrier family 9, isoform A3,	S275	R.EALVEPASES*PRPALAR.S
Solute carrier family 9, isoform A3,	S285*	R.SAS*SDTSEELNSQDSPKR.Q
tankyrase 1 binding protein 1 [Mus	S51	K.DRQS*PSTCSEGLLGWAQK.D
TAO kinase 1	S421*	R.ASDPQS*PPQVSR.H
Transcription factor 12	S583	R.TSS*TNEDEDLNPEQK.I
Transcription factor 20 (AR1)	T683*	R.SLT*PPPSSTESK.V
Transgelin 2	S163	R.NFS*DNQLQEGK.N
Trinucleotide repeat containing	S156	R.S*QSWEERGDR.R
Tumor suppressor p53 binding	S357	K.NQS*SEDILR.D
Vimentin	S56	R.SLYSSS*PGGAYVTR.S
Zinc finger, CW-type with coiled-coil	S679	R.SLAVS*DEEEAEEEAEK.R
Znf HX	S262S266	K.LTPVS*LSNS*PIKGVDCQEVSQEK.N

Figure 5.5: Selected phosphopeptides that were down-regulated in NIH3T3-hIRS1 Y1180F comparing to NIH3T3-hIRS1



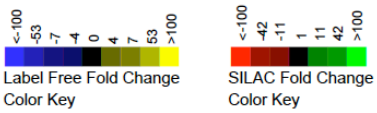
Protein Name	Phosphosites	Sequence
2310047C17Rik protein [Mus	S94	R.KGDRS*PEPGQTWTHEVFSSR.S
5`-3` exoribonuclease 2	S83	R.NSSPS*ISPNTSFASDGSPSPLGGIK.R
53BP1 protein [Mus musculus]	S292	R.ELLEEGPQVQPSS*EPEVSSTQEDLFDQSSK.T
AA589382 protein [Mus musculus]	S921T935	R.S*NSFSDEREFSAPST*PTGTLEFAGGDAK.G
AKT1 substrate 1	T111*	R.LNT*SDFQK.L
Ataxin 2 domain protein	S58*	K.GPPQS*PVFEGVYNNSR.M
Calcium regulated heat stable	S52*	R.TFS*ATVR.A
Catenin, delta 1	S230*	R.HYEDGYPGGSDNYGS*LSR.V
Catenin, delta 1	Y228*	R.HYEDGYPGGSDNY*GSLSR.V
CDC42 effector protein 2	S137S141S145	R.LHLES*PQPS*PQPS*PQGAGNVDVWR.I
CDNA sequence BC022641 [Mus	S860	R.DS*QDTSAEQSDHDDEVASLASASGGFGSK.I
Chromatin accessibility complex,	S122	K.REEEEDNEDDGS*DLGEALA.-
Chromodomain helicase DNA	S1678	R.AASSGPRS*PLDQR.S
Cingulin	S90	R.GSPGALSS*DSELPENPYSQVK.G
Cyclin L1	S379	R.QQASKS*PYNGVR.K
Cytidine 5-prime triphosphate	S575*	R.SGSSS*PDSEITELKFPSISQD.-
Dedicator of cytokinesis 7	S17	R.GQLERS*PSGSAFGSQENLR.W
DKFZP564O123 protein	S199*	K.ATIS*DEEIER.Q
Dynamin 1	T776S778*	R.RSPT*SS*PTPQR.R
Elongation factor 2 kinase	S73*	K.TECGSTGS*PASSFHFK.E
Epidermal growth factor receptor	S658	R.QNS*SSSDSGGSIVR.D
expressed sequence AI116001 [Mus	S617	K.DSQENS*DAELSSSEYIR.A
expressed sequence AI646570 [Mus	S829	R.AES*PETSAVESTQSTPQK.G
expressed sequence AI646570 [Mus	S829	R.AES*PETSAVESTQSTPQK.G
FAK	T741*	K.LQPQEISPPPT*ANLDR.S
Family with sequence similarity 44,	Y480S484	K.Y*YSDS*DDELTVEQR.R
Family with sequence similarity 65,	S22	R.SQS*FAGVLGSHER.G
Filamin B	S2478	R.LVS*PGSANETSSILVESVTR.S
FIP1 like 1	S418	R.ERDHS*PTPSVFNSDEER.Y
Guanine nucleotide binding protein,	S18	R.LEAS*IER.I
hepatoma-derived growth factor	S659	R.TRLAS*ESANDDNEDS.-
Heterochromatin protein 1 alpha	S14	R.TADSSSS*EDEEEYVVEK.V
Heterogeneous nuclear	S284*	R.DYDDMS*PR.R
Heterogenous nuclear	S6*	K.SES*PKEPEQLR.K
IRS 1	S343*	R.ASSDGEGTMSRPASVDGSPVS*PSTNR.T
IRS 1	S1100*	R.HSS*ETFSAPTR.A
Kinesin family member 7	S18	R.S*GSNGSVVSLEQQQK.I
MAP3K7	S432S434	R.SIQDLTVTGTEPGQVSSRSSS*PS*VR.M
mFLJ00139 protein [Mus musculus]	S749	R.KPCPAGSGPSPAALS*PSPSHR.K
Microtubule associated protein 1A	S908	R.CLS*PDDSTVK.M

(Continued from the previous page)



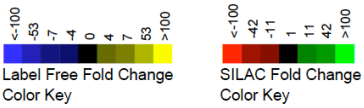
Protein Name	Phosphosites	Sequence
Myristoylated alanine rich protein	S140*	K.AEDGAAPSPS*SETPK.K
NOT2	S242	R.EGS*GNPTPLINPLAGR.A
Nuclear factor kappa B, subunit 2	T425	R.DTDAGEGAEEPRT*PPEAPQGEPQALDTLQR.A
Nuclear valosin containing protein	S190	K.ESLPLDLS*DDQSNSK.K
Oxidation resistance 1	S116	R.VVSSTS*EEEEAFTEK.F
PDGF alpha associated protein 1	S60S63*	K.KSLDS*DES*EDEDDDYQQKR.K
Phosphatidylinositol 4 kinase	S413*	R.S*VENLPECGITHEQR.A
Pleckstrin homology like domain	S450	R.TLQPPES*PR.L
Protein transport protein SEC61	S17*	M.PGPTPSGTNVGSSGRS*PSK.A
RAB3A interacting protein	S218	K.TLVLSSSPTS*PTQEPLAAAK.T
RBMX	S208*	R.DVYLS*PR.D
RIP2	S411	R.AS*SCSLAVISPFLVEK.G
RNA polymerase I and transcript	S42	R.ATEEPSGTGS*DELIK.S
serine/arginine repetitive matrix 2	S972	R.SSS*PVTELTAR.S
serine/arginine repetitive matrix 2	S1402	R.SGSS*QELDGKPSASPQER.S
serine/arginine repetitive matrix 2	S1120	K.SEQPLSQVLPSLS*PEHK.E
serine/arginine repetitive matrix 2	S338	R.HASS*SPESLKPTPAPGSR.R
SET domain containing 2 [Mus	T109	K.T*GSPTPSNQLNDSPTFK.K
Signal induced proliferation	S1544	R.TLS*DESLCSGR.R
Tara like protein	S150	R.KADGPRPSLDYVELSPLAPS*SPQR.M
Thymopoietin	T159*	R.SST*PLPTVSSSAENTR.Q
Topoisomerase DNA II, beta	S1509S1511*	K.IVETINS*DS*DSEFGIPK.K
Transthyretin	T95	R.VELDT*K.S
Tripartite motif containing 47	S277	R.SGALAS*PTDPFQSR.L
Zinc finger protein 740	S44	R.AGS*PDVLR.C
Znf HX	S92	K.YVES*DDEKPTDENVNEK.A

Figure 5.6: Selected phosphopeptides that were up-regulated in NIH3T3-hIRS1 Y1180F comparing to NIH3T3-hIRS1



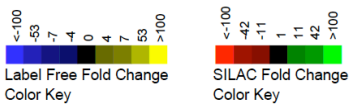
Protein Name	Phosphosites	Sequence
100 kDa thyroid hormone receptor	S840	R.LLSSSDDDANILSS*PTDR.S
100 kDa thyroid hormone receptor	S829	R.LLS*SSDDDANILSSPTDR.S
1810035L17Rik protein [Mus	S104	K.ALHGAQTS*DEER.F
2310047C17Rik protein [Mus	S94	R.KGDRS*PEPGQTWTHEVFSSR.S
3 Phosphoinositide dependent	S241*	R.ANS*FVGTAQYVSPELLTEK.S
5`-3` exoribonuclease 2	S499S501*	R.KAEDS*DS*EPEPEDNVR.L
5`-3` exoribonuclease 2	S448*	R.NS*PGCQVASNPR.Q
5033413A03Rik protein [Mus	S459	R.RQPS*PQPSPR.D
5033413A03Rik protein [Mus	S459S463	R.RQPS*PQPS*PR.D
5033413A03Rik protein [Mus	S459	R.RQPS*PQPSPR.D
53BP1 protein [Mus musculus]	S261	R.SEDRPS*SPQVSVAAVETK.E
53BP1 protein [Mus musculus]	S262	R.SEDRPSS*PQVSVAAVETK.E
6330414O09Rik protein [Mus	S158	R.AGGAS*PAASSTTQPPAQHR.L
AA409316 protein [Mus musculus]	S537	R.KGS*PTPAYPER.K
AA589382 protein [Mus musculus]	S467	K.GDLGASS*PSMK.L
AA589382 protein [Mus musculus]	T318	K.VQANLDT*PDINIEGPEAK.I
AA589382 protein [Mus musculus]	S903	K.ASLGSLEGEVEAEAS*SPKGK.F
AA589382 protein [Mus musculus]	S893	K.ASLGS*LEGEVEAEASSPK.G
AA589382 protein [Mus musculus]	S467	K.GDLGASS*PSMK.L
ACINUS	S477S491	R.S*LSPLSGTTDTKAES*PAGR.V
ACINUS	S81	K.GVQAGNS*DTEGGQPGR.K
Actin binding LIM protein 1	S452*	R.STS*QGSINSPVYSR.H
Actin filament associated protein	S668*	R.SGTSSPQS*PVFR.H
activated c-raf oncogenic fusion	S359S363	R.RGNNS*AVGS*NADLTIEEDEEEEP
activator of G-protein signaling 3	S467	R.APS*SDEECFFDLLSK.F
activity and neurotransmitter-	S55	R.DLYRPLSSDDLDSVGDS*V.-
activity and neurotransmitter-	S51	R.DLYRPLSSDDLDS*VGDSV.-
activity and neurotransmitter-	S45	R.DLYRPLS*SDDLDSVGDSV.-
activity and neurotransmitter-	S46	R.DLYRPLSS*DDLDSVGDSV.-
AF6	S1067	R.TSS*VVTLEVAK.Q
AHNAK nucleoprotein (desmoyokin)	S2801	K.GGVTGS*PEASISGSK.G
AHNAK nucleoprotein isoform 1	S213S217	R.LPSGS*GPAS*PTTGSAVDIR.A
AHNAK nucleoprotein isoform 1	S211	R.LPS*GSGPASPTTGSAVDIR.A
AHNAK nucleoprotein isoform 1	T219	R.LPSGSGPASPT*TGSAVDIR.A
AHNAK nucleoprotein isoform 1	S211T220	R.LPS*GSGPASPTT*GSAVDIR.A
AI840980 protein [Mus musculus]	S65	R.SS*SLGDLLR.E
AI840980 protein [Mus musculus]	S66	R.SSS*LGDLLR.E
AKT1 substrate 1	S68S76*	R.TEARSS*DEENGPPS*SPDLDR.I
AKT1 substrate 1	S67S76*	R.TEARS*SDEENGPPS*SPDLDR.I
AKT1 substrate 1	S67S77*	R.TEARS*SDEENGPPSS*PDLDR.I
Amphiphysin II	S304*	K.SPSPPPDGS*PAATPEIR.V
Angiomotin like 2	S443	R.TPS*LDSIAATR.V

(Continued from the previous page)



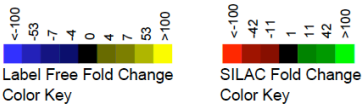
Protein Name	Phosphosites	Sequence
Angiomotin like 2	S443	R.TPS*LDSIAATR.V
Ankyrin repeat domain-containing	S82	R.FCTGDS*PPLEAK.L
annexin V-binding protein ABP-7	S14	K.TSFDENDS*EELEDKDSK.S
AP2 associated kinase 1	T620S624*	K.VGSLT*PPSS*PK.T
AP2 associated kinase 1	S637	R.ILS*DVTHSAVFGVPASK.S
AP2 associated kinase 1 isoform 2	T523	K.VQTT*PPPTIQGQK.V
ATP binding cassette 50	S107*	K.QLSVPAS*DEEDEVPAPIPR.G
ATP-binding domain-containing	S266	K.VSSVPADDETANSIHS*S.-
Atrophin 1	S77*	R.SEEIS*ESESEETSAPK.K
B-Raf	S291*	R.SAS*EPSLNR.A
BC003993 protein [Mus musculus]	S550	R.RDS*FSENEKQR.S
BCL2 associated transcription factor	S654*	R.IDIS*PSALR.K
BCL2 associated transcription factor	S175*	K.KAEGEPQEES*PLKSK.S
BIP	S650	K.LYGSGGPPPTGEEDTS*EKDEL.-
BRG1 associated factor 180 KD	S1453	R.AAQQQQPSAS*PR.A
Bystin	S97	R.LGPGLPQDGS*DEEDEEWPTLEK.A
C11orf23 protein	S537*	R.IQQFDDGGS*DEEDIWEEK.H
C14orf4 protein	S657*	R.NS*SSPVSPASVPGQR.R
Cancer susceptibility candidate 3	S145	K.GTVTGERQS*GDGQESTEPVENK.V
Casein kinase 1, delta	S382	R.GAPVNVS*SSDLTGR.Q
CASK [Homo sapiens]	S571	R.TQSSS*CEDLPSTTQPK.G
Catechol-O-methyltransferase	S261	K.AVYQGPGSS*PVKS.-
Catenin alpha 1	S641*	R.TPEELDDS*DFETEDFDVR.S
CDNA sequence BC022641 [Mus	S864	R.DSQDTS*AEQSDHDDEVASLASASGGF
cDNA sequence BC037996;	S493	R.ELS*PEQSTAGKPSDGSSALDR.A
Cell division cycle 2 like 5	S383*	R.GGDVS*PSPYSSSSWR.R
Chromodomain helicase DNA	T1111	R.IDGGIT*GALR.Q
Clk3 protein [Mus musculus]	S131	K.SQDVAIS*PQQQQCSK.S
Coiled-coil domain-containing	S165	R.LKGQEDSLASAVDATTGQEACDS*D.-
Conserved nuclear protein NHN1	S299*	K.LGVSVS*PSR.A
cyclin Y-like 1 [Mus musculus]	S276	R.SLS*ADNFIGIQR.S

(Continued from the previous page)



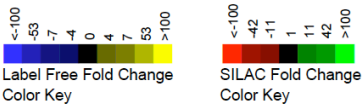
Protein Name	Phosphosites	Sequence
DDX10	T587	R.LASGDGDEEQDEET*EDEETEDHLGK.A
DDX24	S80*	K.AQAVS*EEEEEEEGQSSSPK.K
DEAD-box protein 54	S74	R.ALPSFPTSECVS*DVEPDTR.E
Death associated protein	S51	R.DKDDQEWESTS*PPKPTVFISGVIAR.G
DNA segment, Chr 6, Wayne State	S873	R.SSVPSGGS*LFGDDEDDDLFSSAK.T
Dynein light chain A	S510	R.KPASVS*PTTPTSPTEGEAS.-
DYRK1B	Y37*	R.IYQY*IQSR.F
Dyskerin	S481	K.TVLES*GGETGDGDNDTTK.K
eIF 4G1	S1146	R.S*FSKEVEER.S
eIF 4G1	S1148	R.SFS*KEVEER.S
eIF-2B epsilon	S539	R.AGS*PQLDDIR.V
Enabled homolog	T308	K.APSTST*PEPTR.K
ERCC5	S384*	R.NSGATADAGSIS*PR.T
eukaryotic translation elongation	S106	K.YGPSSVEDTTGSGAADAKDDDDIDLFGS*DDEEES
Eukaryotic translation elongation	S162	K.GK.AQDATTGVLWNEEDT*DEDEDDDGVSPESHPF
Eukaryotic translation elongation	S119	R.LSSLEKSS*PTPR.A
eukaryotic translation initiation factor	S120S123	R.GHPSAGAEEEGGS*DGS*AAEAEPR.A
Eukaryotic translation initiation factor	S951	R.SSS*PVGLAK.W
expressed sequence AA591047	S665	R.SSS*VGSSSSYPISSAGPR.T
F box protein 4	S11*	R.GAGS*PPPASDWGR.L
Filamin C	S2213	R.LGS*FGSITR.Q
Forkhead box J3	S223	K.VTLYNADQDGSDS*PR.S
G patch domain-containing protein 4	S258	R.QGTAIGS*EEEEAAGESGPR.E
GFAT	S243	R.VDS*TTCLFPVEEK.A
GFAT	T245	R.VDSTT*CLFPVEEK.A
GIP90	S1049	R.SNS*NSSSVITTEDNK.I
Glucocorticoid receptor DNA binding	S1159*	R.TSFSVGS*DDELGPIR.K
Glycogen synthase kinase 3 beta	Y279*	R.GEPNVSY*ICSR.Y
GTPase activating Rap/RanGAP	S241	R.SSS*TSDILEPFTVER.A
Heat shock protein 105 kDa	S810*	K.IES*PKLER.T
Hepatoma derived growth factor	S165*	R.AGDVLEDS*PKRPK.E
Heterogeneous nuclear	S359	R.SSGS*PYGGGYGSGGGSGGYGSR.R
High mobility group AT hook 1	S102	K.KLEKEEEEGISQES*SEEEQ.-
High mobility group AT hook 2	S100S104	K.KPAQETEETS*SQES*AEED.-
HLA-B associated transcript 3	S995*	R.TSPEPQRENAS*PAPGTTAEEAMSR.G
HSP90B	S255*	K.IEDVGS*DEEDDSGKDK.K
HSP90B	S255*	K.IEDVGS*DEEDDSGKDKK.K

(Continued from the previous page)



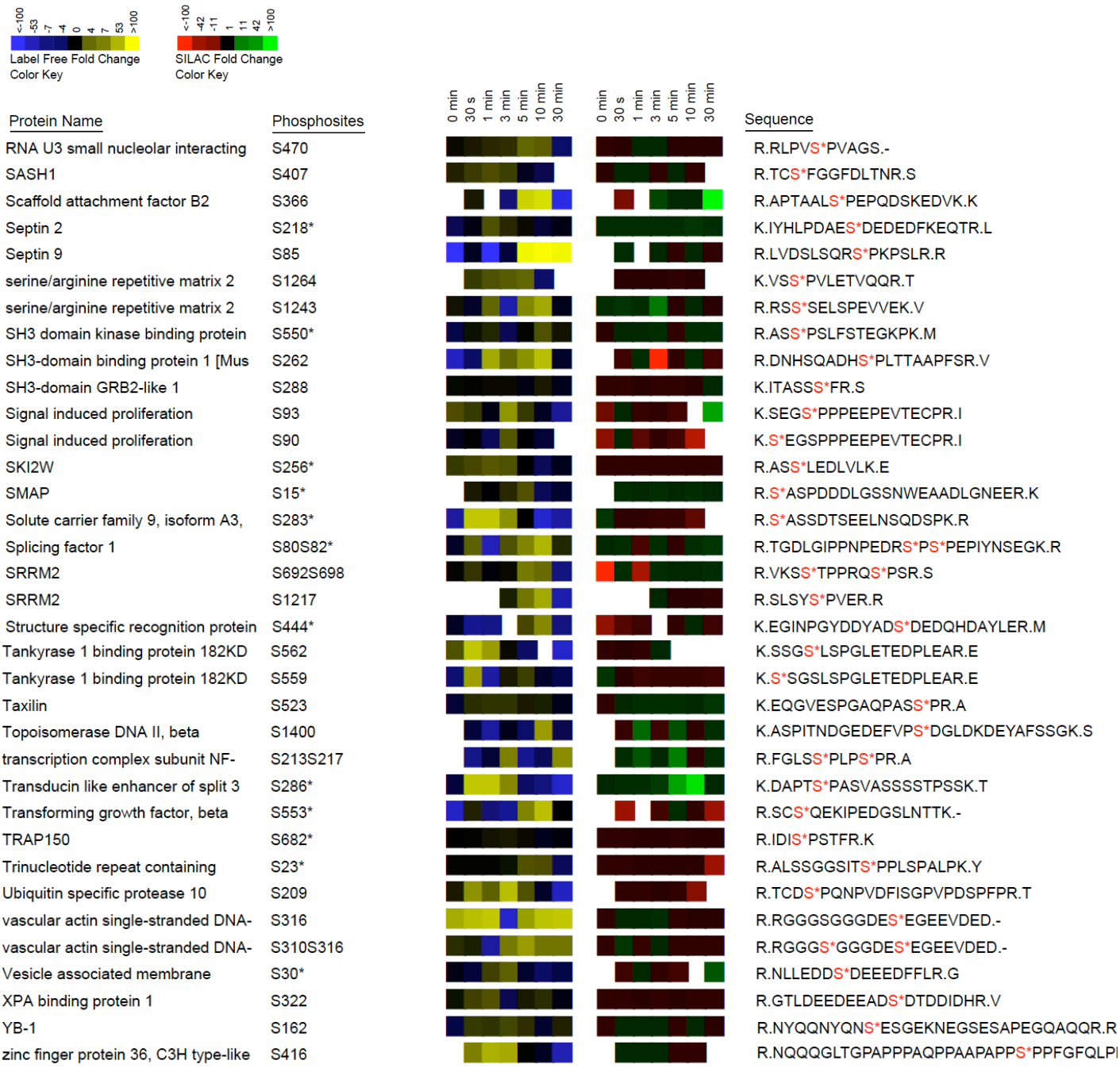
Protein Name	Phosphosites	Sequence
IGF II mRNA binding protein 1	S181*	R.QGS*PVAAGAPAK.Q
IGF II mRNA binding protein 3 [Mus	S165	K.VAYIPDETAAQQNPS*PQLR.G
IRS 1	S1077	R.VNLS*PNHNQSAK.V
IRS2	T517*	R.SNT*PESIAETPPAR.D
Karyopherin alpha3	S60	R.NVPQEESLEDS*DVDADFK.A
Kinesin family member 21B	S1217	R.GATDTS*PLTR.R
Lamin A/C	S632	R.SVGGS*GGGSFGDNLVTR.S
LATS1	S296*	R.SNS*FNNPLGSR.A
LRRFIP2	S320	R.NSASATTPLS*GNSSR.R
MAP3K7	S412	R.RRS*IQDLTVTGTEPGQVSSR.S
MARCKS like protein	T148*	K.AAAT*PESQEPQAK.G
Matrin 3	S619	K.TDAQKTES*PAEGKEQEEK.S
mBLVR [Mus musculus]	S760	R.HSSLPTES*DEDIAPAQR.V
MCM2 minichromosome	S27*	R.ISDPLTSS*PGR.S
MCM2 minichromosome	S139*	R.GLLYDS*SEEDEERPAR.K
MCM2 minichromosome	S139*	R.RGLLYDS*SEEDEERPAR.K
MCM2 minichromosome	S139S140*	R.GLLYDS*S*EEDEERPAR.K
MCM3	S655	K.ASEDES*DLEDEEEKSQEDTEQKR.K
Microtubule associated protein 1B	S1293S1307	R.SVS*PGVTQAVVEEHCAS*PEEK.T
Microtubule associated protein 4	S785	K.ATS*PSTLVSTGPSSR.S
Mitogen-activated protein kinase	S340*	R.SADS*ENALTVQER.N
Myc-induced SUN domain	S23	R.FQQPPQPEGEEDAS*DGGR.K
Myosin heavy chain 9, nonmuscle	S1944*	K.GTGDCS*DEEVDGK.A
Myosin heavy chain 9, nonmuscle	S1944*	K.GTGDCS*DEEVDGKADGADAK.A
Myosin heavy chain 9, nonmuscle	S1944*	R.KGTGDCS*DEEVDGKADGADAK.A
MYST histone acetyltransferase 2	S57	R.LSQSSQDSS*PVR.N
NICE 4 protein	S470*	K.STSAPQMSPGS*SDNQSSSPQPAQQK.L
NICE 4 protein	S608*	R.YPSSISS*SPQKDLTQAK.N
NimA-related protein kinase 3	T475*	R.LDEEDT*DFEEDNENPDWVSELKK.H
NSFL1C	S272	K.LGSTAPQVLNTSS*PAQQAENEAK.A
Nuclear factor I/X	S265	R.S*ITSPPSTSSTK.R
Nuclear ubiquitous casein kinase	S19*	R.KVVDYSQFQES*DDADEDYGR.D
nucleolar protein p120 - mouse	T130	K.AQDATTGVLWNEEDT*DEDEDDDGVSPES
ortholog of human Ras association	S192	R.TAS*AGTVSDAEAR.S
OSBP related protein 11	S186*	R.SFSLASSGNS*PISQR.R

(Continued from the previous page)



Protein Name	Phosphosites	Sequence
P160	S1280	K.LSQVNGATPVS*PIEPESK.K
PAXIP1-associated protein 1 (PTIP-	S236	R.DLFSLDSEGPSPTS*PPLR.S
PDGF alpha associated protein 1	S57S63*	K.KS*LDSDES*EDEDDDYQQK.R
PDGF alpha associated protein 1	S60S63*	K.KSLDS*DES*EDEDDDYQQK.R
PDZ and LIM domain 2	T292	R.ALAT*PPKLHTCEK.C
pericentriolar material-1 [Mus	S65	R.VTNAIS*PESSPGVGR.R
phosphatidylinositol 4-kinase type II-	S462	R.SAS*ESYTQSFQSR.K
Phosphatidylinositol-3-	S318	R.SAS*ITNLSLDR.S
phosphoprotein	T925	K.GGEFDEFVNDDT*DDDLPVSK.K
phosphoprotein	S941S943	R.KGS*GS*EQEGEEEEGGER.K
Pleckstrin homology like domain	S468	R.ELPPLS*PSLSR.R
Plectin 1	S21	K.RTSS*EDNLYLAVLR.A
plenty-of-prolines-101; POP101;	S714	R.APQTSS*PPPVR.R
Polypyrimidine tract binding protein	S141	K.ELKTDSS*PNQAR.A
Programmed cell death 5	S119	R.KVMDS*DEDDADY.-
Protein kinase C and casein kinase	S354	R.DGTAPPPQSPSSPGSGQDEDWS*DEESPR
Protein kinase, cAMP dependent,	S83*	R.TDSREDEIS*PPPPNPVVK.G
Protein kinase, cAMP dependent,	S77*	R.TDS*REDEISPPPPNPVVK.G
Protein kinase, cAMP dependent,	S114*	R.RAS*VCAEAYNPDEEEDDAESR.I
protein phosphatase 2 regulatory	S81	R.QSS*FPFNLNK.N
Protein tyrosine phosphatase, non	S314	K.ICTEQSNS*PPPIR.R
PtdIns 4-kinase [Homo sapiens]	S294	R.TAS*NPKVENEDEPVR.L
PTPRF interacting protein alpha 1	S668	R.VGSGS*LDNLGR.F
PWP2	S895S898	R.TLEPVDTEEDS*DAS*DEDSLHLLR.A
PYGO2	T359	R.GGGT*PDANSLAPPGK.A
rA1	S829	K.GTEETSWS*GEER.T
Rabaptin 5	S407S410	R.AQS*TDS*LGTSSSLQSK.A
Ran binding protein 3	S99	R.VLS*PPKLNEANSDTSR.E
Ran-binding protein 3 (RanBP3)	S58	R.TSS*LTHSEEK.S
Ras-GTPase-activating protein SH3-	S231*	K.STS*PAPADVAPAQEDLR.T
RBMX	S208*	R.DVYLS*PRDDGYSTK.D
Rho guanine nucleotide exchange	S906	K.LTSVLS*PR.L
Rho guanine nucleotide exchange	S151*	K.SVS*TTNIAGHFNDESPLGLR.Q
RNA binding motif protein 15	S294*	R.SLS*PGGAALGYR.D
RNA binding motif protein 25	S107*	K.LGASNS*PGQPNSVK.R

(Continued from the previous page)



Protein Name	Phosphosites	Sequence
RNA U3 small nucleolar interacting	S470	R.RLPVS*PVAGS.-
SASH1	S407	R.TCS*FGGFDLTNR.S
Scaffold attachment factor B2	S366	R.APTAALS*PEPQDSKEDVK.K
Septin 2	S218*	K.IYHLPDAES*DEDEDFKEQTR.L
Septin 9	S85	R.LVDSLSQRS*PKPSLR.R
serine/arginine repetitive matrix 2	S1264	K.VSS*PVLETVQQR.T
serine/arginine repetitive matrix 2	S1243	R.RSS*SELSPEVVEK.V
SH3 domain kinase binding protein	S550*	R.ASS*PSLFSTEGKPK.M
SH3-domain binding protein 1 [Mus	S262	R.DNHSQADHS*PLTTAAPFSR.V
SH3-domain GRB2-like 1	S288	K.ITASSS*FR.S
Signal induced proliferation	S93	K.SEGS*PPPEEPEVTECPR.I
Signal induced proliferation	S90	K.S*EGSPPPEEPEVTECPR.I
SKI2W	S256*	R.ASS*LEDLVLK.E
SMAP	S15*	R.S*ASPDDDLGSSNWEAADLGNEER.K
Solute carrier family 9, isoform A3,	S283*	R.S*ASSDTSEELNSQDSPK.R
Splicing factor 1	S80S82*	R.TGDLGIPPNPEDRS*PS*PEPIYNSEGK.R
SRRM2	S692S698	R.VKSS*TPPRQS*PSR.S
SRRM2	S1217	R.SLSYS*PVER.R
Structure specific recognition protein	S444*	K.EGINPGYDDYADS*DEDQHDAYLER.M
Tankyrase 1 binding protein 182KD	S562	K.SSGS*LSPGLETEDPLEAR.E
Tankyrase 1 binding protein 182KD	S559	K.S*SGSLSPGLETEDPLEAR.E
Taxilin	S523	K.EQGVESPGAQPASS*PR.A
Topoisomerase DNA II, beta	S1400	K.ASPITNDGEDEFVPS*DGLDKDEYAFSSGK.S
transcription complex subunit NF-	S213S217	R.FGLSS*PLPS*PR.A
Transducin like enhancer of split 3	S286*	K.DAPTS*PASVASSSSTPSSK.T
Transforming growth factor, beta	S553*	R.SCS*QEKIPEDGSLNTTK.-
TRAP150	S682*	R.IDIS*PSTFR.K
Trinucleotide repeat containing	S23*	R.ALSSGGSITS*PPLSPALPK.Y
Ubiquitin specific protease 10	S209	R.TCDS*PQNPVDFISGPVPDSPFPR.T
vascular actin single-stranded DNA-	S316	R.RGGGSGGGDES*EGEEVDED.-
vascular actin single-stranded DNA-	S310S316	R.RGGGS*GGGDES*EGEEVDED.-
Vesicle associated membrane	S30*	R.NLLEDDS*DEEEDFFLR.G
XPA binding protein 1	S322	R.GTLDEEDEEADS*DTDDIDHR.V
YB-1	S162	R.NYQQNYQNS*ESGEKNEGSESAPEGQAQQR.R
zinc finger protein 36, C3H type-like	S416	R.NQQQGLTGPAPPPAQPPAAPAPPS*PPFGFQLPI

Figure 5.7: Selected phosphopeptides that showed no obvious change between NIH3T3-hIRS1 Y1180F and NIH3T3-hIRS1

5.4 CONCLUSION

In this project we aimed to use our novel high-throughput proteomic pipeline to explore the change in phosphorylation profiles in different NIH3T3 cells in response to insulin stimulation. In a SILAC-labeled NIH3T3-hIRS1/NIH3T3-hIRS1 Y1180F timecourse, we discovered a total of 2201 phosphorylation sites at an estimated 1% false discovery rate. While previous studies have shown that SHP2 has a positive effect on MAPK/Ras cascade in insulin signaling pathway but has little influence on PI3K/Akt cascade [32, 33], we discovered hundreds of phosphorylation sites that were up-regulated, down-regulated and were unaffected in the NIH3T3-hIRS1 Y1180F cells (hIRS-1 without SHP2 binding site). Although some of the unchanged sites were related to routine cell activities, some of them were clearly involved in the PI3K/Akt branch.

Using protein-protein interaction exploration software developed in the lab of Prof. David Laidlaw at Brown, we will explore the connections between quantitatively perturbed phosphorylation sites and expected protein interactions. Future experiments will examine selected protein-protein interactions of proteins discovered in this experiment showing significant changes between NIH3T3-hIRS1 Y1180F and NIH3T3-hIRS1 with a particular emphasis on canonical insulin signaling pathway proteins. Data generated in this experiment is useful to study binding substrates of a particular phosphoprotein by co-immunoprecipitation, expressing TAG and pull down, or synthesizing phosphopeptides discovered here and applying pull-down assay.

5.5 REFENRENCE

1. Myers, M.G., Jr. and M.F. White, *Insulin signal transduction and the IRS proteins.* Annu Rev Pharmacol Toxicol, 1996. **36**: p. 615-58.

2. Czech, M.P., *The nature and regulation of the insulin receptor: structure and function.* Annu Rev Physiol, 1985. **47**: p. 357-81.

3. Myers, M.G., Jr., X.J. Sun, and M.F. White, *The IRS-1 signaling system.* Trends Biochem Sci, 1994. **19**(7): p. 289-93.

4. White, M.F. and C.R. Kahn, *The insulin signaling system.* J Biol Chem, 1994. **269**(1): p. 1-4.

5. Myers, M.G., Jr., et al., *IRS-1 activates phosphatidylinositol 3'-kinase by associating with src homology 2 domains of p85.* Proc Natl Acad Sci U S A, 1992. **89**(21): p. 10350-4.

6. Backer, J.M., et al., *Phosphatidylinositol 3'-kinase is activated by association with IRS-1 during insulin stimulation.* EMBO J, 1992. **11**(9): p. 3469-79.

7. Kitamura, T., et al., *Insulin-induced phosphorylation and activation of cyclic nucleotide phosphodiesterase 3B by the serine-threonine kinase Akt.* Mol Cell Biol, 1999. **19**(9): p. 6286-96.

8. Cross, D.A., et al., *Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B.* Nature, 1995. **378**(6559): p. 785-9.

9. Datta, S.R., et al., *Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery.* Cell, 1997. **91**(2): p. 231-41.

10. Brunet, A., et al., *Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor.* Cell, 1999. **96**(6): p. 857-68.

11. Andjelkovic, M., et al., *Role of translocation in the activation and function of protein kinase B.* J Biol Chem, 1997. **272**(50): p. 31515-24.

12. Xiao, S., et al., *Syp (SH-PTP2) is a positive mediator of growth factor-stimulated mitogenic signal transduction.* J Biol Chem, 1994. **269**(33): p. 21244-8.

13. Baltensperger, K., et al., *Binding of the Ras activator son of sevenless to insulin receptor substrate-1 signaling complexes.* Science, 1993. **260**(5116): p. 1950-2.

14. Jhun, B.H., et al., *Insulin and insulin-like growth factor-I signal transduction requires p21ras.* J Biol Chem, 1994. **269**(8): p. 5699-704.

15. Kuhne, M.R., et al., *The insulin receptor substrate 1 associates with the SH2-containing phosphotyrosine phosphatase Syp.* J Biol Chem, 1993. **268**(16): p. 11479-81.

16. Skolnik, E.Y., et al., *The function of GRB2 in linking the insulin receptor to Ras signaling pathways.* Science, 1993. **260**(5116): p. 1953-5.

17. Yonezawa, K., et al., *Signal transduction pathways from insulin receptors to Ras. Analysis by mutant insulin receptors.* J Biol Chem, 1994. **269**(6): p. 4634-40.

18. Tanaka, S., T. Ito, and J.R. Wands, *Neoplastic transformation induced by insulin receptor substrate-1 overexpression requires an interaction with both Grb2 and Syp signaling molecules.* J Biol Chem, 1996. **271**(24): p. 14610-6.

19. Ogawa, W., T. Matozaki, and M. Kasuga, *Role of binding proteins to IRS-1 in insulin signalling.* Mol Cell Biochem, 1998. **182**(1-2): p. 13-22.

20. Nishiyama, M. and J.R. Wands, *Cloning and increased expression of an insulin receptor substrate-1-like gene in human hepatocellular carcinoma.* Biochem Biophys Res Commun, 1992. **183**(1): p. 280-5.

21. Ito, T., Y. Sasaki, and J.R. Wands, *Overexpression of human insulin receptor substrate 1 induces cellular transformation with activation of mitogen-activated protein kinases.* Mol Cell Biol, 1996. **16**(3): p. 943-51.

22. de la Monte, S.M., et al., *Differential effects of ethanol on insulin-signaling through the insulin receptor substrate-1.* Alcohol Clin Exp Res, 1999. **23**(5): p. 770-7.

23. Nguyen, V., et al., *A new approach for quantitative phosphoproteomic dissection of signaling pathways applied to T cell receptor activation.* Mol. Cell. Proteomics, 2009. **in press**.

24. Larsen, M.R., et al., *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns.* Mol Cell Proteomics, 2005. **4**(7): p. 873-86.

25. Ficarro, S.B., et al., *Automated immobilized metal affinity chromatography/nano-liquid chromatography/electrospray ionization mass spectrometry platform for profiling protein phosphorylation sites.* Rapid Commun Mass Spectrom, 2005. **19**(1): p. 57-71.

26. Licklider, L.J., et al., *Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column.* Anal Chem, 2002. **74**(13): p. 3076-83.

27. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

28. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

29. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nat Biotechnol, 2006. **24**(10): p. 1285-92.

30. Hunter, T. and B.M. Sefton, *Transforming gene product of Rous sarcoma virus phosphorylates tyrosine.* Proc Natl Acad Sci U S A, 1980. **77**(3): p. 1311-5.

31. Sun, X.J., et al., *Pleiotropic insulin signals are engaged by multisite phosphorylation of IRS-1.* Mol Cell Biol, 1993. **13**(12): p. 7418-28.

32. Yamauchi, K., et al., *Protein-tyrosine-phosphatase SHPTP2 is a required positive effector for insulin downstream signaling.* Proc Natl Acad Sci U S A, 1995. **92**(3): p. 664-8.

33. Chen, H., et al., *Protein-tyrosine phosphatases PTP1B and syp are modulators of insulin-stimulated translocation of GLUT4 in transfected rat adipose cells.* J Biol Chem, 1997. **272**(12): p. 8026-31.

Chapter 6

CONCLUSIONS

6.1 SUMMARY OF RESULTS

Since biological mass spectrometry has been introduced in 1990s, exciting progress in the development of new instrumentation and high throughput proteomic methods has led to a landslide of proteomic data that needs to be analyzed and explored efficiently. However, this enhanced ability in data acquisition has not been accompanied by a concomitant increase in the availability of flexible tools that allows users to rapidly assimilate, explore, and analyze this data and adapt to a variety of experimental workflows with minimal user intervention. Often the manual aggregation of proteomic data and analysis in current proteomics software distract investigators from the biological meaning of their data, leading to the all-too-frequent deposition of proteomic data into the scientific literature with little or no biological or clinical interpretation. We fill the critical gap by providing a high-throughput autonomous proteomic pipeline to streamline the total analysis of a complex proteomic sample. The work presented in this thesis covered the most critical components within the proteomic pipeline.

High-throughput proteomic methods necessitate the development of large-scale statistical spectral validation algorithms. We developed a novel method based on a logistic regression model that learns rules from a user-provided training set. Our statistical approach mimics experts' manual spectra validation criteria and provides an estimated spectral confidence level. The result showed a substantial improvement for this method over the generic score provided by database search engine, such as XCorr by Sequest and E-value by X!Tandem. A flexible platform was developed to tie multiple pieces of proteomic software together to streamline the data processing. Our platform was capable of performing LC/MS acquisition control, MS/MS database search, peptide

spectral validation, phosphorylation site localization and peptide quantitation. The resulting peptide identifications, along with data-dependent calculation results were directed into a relational FileMaker/MySQL database for organization of expansive proteomic data sets, collation of proteomic data with available protein information resources, and visual comparison of multiple quantitative proteomic experiments. An information-rich user interface was presented to end-users to unravel the biological significance of acquired proteomic data.

We also explored the utility of our bioinformatic tools in the analysis of insulin signaling in hepatocellular carcinoma. Using a hybrid quantitation approach combining label-free and SILAC, we were able to quantify a total of 2201 phosphorylation sites at a 1% false discovery rate. Several categories of phosphorylation pattern were identified based on the quantitative data.

6.2 FUTURE WORK

A major question that remains unanswered is how to use the collected quantitative phosphoproteomic timecourse data to predict signal transduction events happened during cell activation. Due to the inherent complexity of signaling networks, machine learning or statistical approaches must be employed. Comparative data generated by mutated signaling molecules may provide a hint in the placement of phosphorylated proteins in the signaling network. Furthermore, new text mining tools may be developed to reveal the internal connections between acquired datasets with existing protein and network knowledge contained in the literature. The future direction of bioinformatic software for

phosphoproteomic analysis should focus on understanding the protein-protein interaction networks assembled from the quantitative timecourse data.