ALGORITHMS FOR SHOTGUN PROTEOMICS SPECTRAL IDENTIFICATION AND QUALITY ASSESSMENT

By

Ze-Qiang Ma

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

In

Biomedical Informatics

May, 2012

Nashville, Tennessee

Approved:

Professor David L. Tabb

Professor Daniel C. Liebler

Professor Bing Zhang

Professor Kathleen L. Gould

Professor Zhongming Zhao

ACKNOWLEDGMENTS

I would like to express profound gratitude to my advisor, Dr. David L. Tabb, for his invaluable support, supervision and helpful suggestions throughout all my graduate school research work. I am also grateful to my other dissertation committee members, Dr. Daniel C. Liebler, Dr. Bing Zhang, Dr. Kathleen L. Gould and Dr. Zhongming Zhao, who were very supportive of my research and provided valuable advice on my dissertation work.

I would like to thank other members in Tabb group, particularly Dr. Surendra Dasari and our star programmer Matt Chambers for their tremendous help in my research. I found it always fun to work with them and I learn something new every day from them. I am also grateful to Dr. Amy-Joan L. Ham, Dr. Stacy D. Sherrod and Dr. Robbert Slebos at the Jim Ayers Institute for Precancer Detection and Diagnosis at Vanderbilt University for providing testing data sets and helpful discussions for my dissertation work.

Finally, I would like to express my gratitude to my wife Yang Wang and our lovely daughter Olivia Ma for all unconditional supports and patience. I want to thank my parents for being ever so understanding and supportive.

Thanks to NIH grants R01 CA126218 and U24 CA126479 for supporting my research work.

ABBREVIATIONS

1D, 2D	One-Dimensional, Two-Dimensional
BSA	Bovine Serum Albumin
CID	Collision Induced Dissociation
CPTAC	Clinical Proteomic Tumor Analysis Consortium
Da	Dalton
DNA	DeoxyriboNucleic Acid
DTT	DiThioThreitol
ESI	ElectroSpray Ionization
ETD	Electron Transfer Dissociation
FDR	False Discovery Rate
FPR	False Positive Rate
FTICR	Fourier Transform Ion Cyclotron Resonance
GUI	Graphical User Interface
HCD	Higher-energy Collision Dissociation
HPLC	High Pressure Liquid Chromatography
ID	IDentification
IMAC	Immobilized Metal Ion Affinity Chromatography
MALDI	Matrix Assisted Laser Desorption and Ionization
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry

MudPIT	Multidimensional Protein Identification Technology
NCI	National Cancer Institute
NGS	Next Generation Sequencing
NIST	National Institute of Standards and Technology
OMSSA	Open Mass Spectrometry Search Algorithm
PEP	Posterior Error Probability
ppm	parts per million
PSM	Peptide Spectrum Match
PTM	Post-Translational Modification
QC	Quality Control
RNA	RiboNucleic Acid
ROC	Receiver Operating Characteristic
RP	Reverse Phase
RP-HPLC	Reverse Phase High Pressure Liquid Chromatography
SCX	Strong Cation Exchange
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
S/N	Signal-to-Noise ratio
TCGA	The Cancer Genome Altas
XIC	Extracted Ion Chromatograms
TOF	Time Of Flight

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

ABBREVIATIONS iii

LIST OF TABLES ix

LIST OF FIGURES x

Chapter

I. INTRODUCTION 1

I.1 Mass Spectrometry-Based Proteomics 1

I.1.1 Overview 1

I.1.2 Sample Preparation and Separation 3

I.1.3 Protein Digestion 4

I.1.4 Mass Spectrometry Instruments 5

I.1.5 Peptide Fragmentation 7

I.2 Proteomics Data Analysis 11

I.2.1 Overview 11

I.2.2 Peptide Identification 14

I.2.3 Peptide Validation 20

I.2.4 Protein Inference 27

I.3 Instrumentation Quality Control 31

I.4 Dissertation Outline 32

II. IDBOOST: VALIDATION AND RESCUE OF TANDEM MASS SPECTRAL IDENTIFICATIONS VIA SPECTRAL CLUSTERING 33
II.1 Introduction 33
II.2 Algorithm 35
II.2.1 Overview 35
II.2.2 Spectral Clustering 36
II.2.3 Rescue of Spectral Identifications 37
II.2.4 Bayesian Average Score 39
II.3 Data Sources 41
II.4 Results and Discussion 44
II.4.1 Rescue of Phosphopeptide Spectra to Resolve Phosphosite Localization Ambiguity 45
II.4.2 Rescue of Spectra in Comparative Analysis 49
II.4.3 Rescue of Spectra in a Variety of Datasets 52
II.5 Conclusion 54

III. SCANRANKER: QUALITY ASSESSMENT OF TANDEM MASS SPECTRA VIA SEQUENCE TAGGING 56
III.1 Introduction 56
III.2 Algorithm 59
III.2.1 Overview 59
III.2.2 BestTagScore Subscore 62
III.2.3 BestTagTIC Subscore 62
III.2.4 TagMzRange Subscore 63

III.2.5 Spectral Quality Score 63
III.3 Data Sources 64
III.4 Results and Discussion 73
III.4.1 Subscore Evaluation 74
III.4.2 Removal of Low Quality Spectra 75
III.4.3 Recovery of Unidentified High Quality Spectra 78
III.4.4 Comparison of ScanRanker to QualScore 80
III.4.5 Prediction of Richness of Identifiable Spectra 81
III.4.6 Use of Quality Score in Peptide Validation 83
III.4.7 Selection of Spectra for *De Novo* Sequencing 85
III.4.8 Use of ScanRanker in Cross-linking Analysis 87
III.5 Conclusion 89

IV. QUAMETER: MULTI-VENDOR PERFORMANCE METRICS FOR LC-MS/MS PROTEOMICS INSTRUMENTATION 91
IV.1 Introduction 91
IV.2 Overview 92
IV.3 Data Sources 94
IV.4 Results and Discussion 100
IV.4.1 Differences between QuaMeter and MSQC 100
IV.4.2 Multi-vendor Performance 103
IV.4.3 Impact of identification tools 106
IV.5 Conclusion 108

V. DISCUSSION 110

V.1 Summary of Results 110

V.2 Future Direction 112

V.2.1 Peptide Identification 112

V.2.2 PTM Identification and Validation 114

V.2.3 Next Generation Sequencing and Proteomics 114

V.2.4 Integration of Omics Data 115

V.2.5 Targeted Proteomics 116

Appendix

A. SOFTWARE CONFIGURATIONS 117

MyriMatch Configurations 117

Sequest Configurations 119

X!Tandem Configurations 119

PepNovo Configurations 120

TagRecon Configurations 120

Pepitome Configurations 121

ScanRanker Configurations 121

IDPicker Configurations 121

QuaMeter Configurations 121

REFERENCES 123

LIST OF TABLES

Table Page

Table 1. Bioinformatics tools for MS-based proteomics data analysis. 13

Table 2. Experimental datasets for the evaluation of IDBoost. 42

Table 3. Experimental datasets for the evaluation of ScanRanker. 66

Table 4. Experimental datasets for the evaluation of QuaMeter. 94

LIST OF FIGURES

Figure Page

Figure 1. The typical MS-based proteomics workflow. 2

Figure 2. Theoretical fragmentation of a peptide..... 8

Figure 3. Mobile proton model for peptide fragmentation. 10

Figure 4. The typical MS-based proteomics data analysis workflow. 12

Figure 5. Four peptide identification strategies. 14

Figure 6. Peptide identification by the database search strategy. 15

Figure 7. Score distribution for correct and incorrect PSMs. 23

Figure 8. A simplified example of protein inference. 28

Figure 9. A diagram of rescuing unidentified spectra in a cluster. 38

Figure 10. Analysis of rescued PSMs in phosphorylation studies. 47

Figure 11. Impact of IDBoost on recognition of differentially expressed proteins in comparative analysis. 51

Figure 12. IDBoost performance in a variety of datasets. 54

Figure 13. A screenshot of ScanRanker GUI. 60

Figure 14. A screenshot of IonMatcher GUI. 62

Figure 15. Combining three subscores improves the discriminating power of ScanRanker. 74

Figure 16. Removing poor MS/MS scans in ScanRanker does not significantly reduce identifications. 76

Figure 17. Determine spectral removal threshold from a single replicate. 77
Figure 18. Evaluation of ScanRanker to recover unidentified high quality spectra. 79
Figure 19. Comparison of ScanRanker to QualScore. 81
Figure 20. ScanRanker scores predict the richness of identifiable spectra. 83
Figure 21. Adding ScanRanker scores in peptide validation increases the number of confident spectrum identifications. 84
Figure 22. ScanRanker scores can be used to predict *de novo* sequencing success. 86
Figure 23. ScanRanker helps to prioritize spectra for manual inspection in cross-linking analysis. 89
Figure 24. Workflow diagram for QuaMeter operation 93
Figure 25. QuaMeter generates similar metrics as MSQC except several chromatographic metrics due to the use of distinct chromatogram extraction tools. 101
Figure 26. QuaMeter generates reliable chromatographic data in instruments from multiple vendors via the Crawdad function in ProteoWizard. 102
Figure 27. QuaMeter computes QC metrics for multiple instrument platforms. 104
Figure 28. QuaMeter metrics help to spot abnormal instrument performance. 106
Figure 29. Distinct identification tools produce different QC metrics with similar variation. 108
Figure 30. A summary of three bioinformatics tools in proteomics data analysis workflow. 111

CHAPTER I

INTRODUCTION

The topic of this dissertation is the development of novel algorithms and bioinformatics tools for proteomics data analysis. This chapter provides a general introduction to the field of proteomics and the data analysis process. The following is not intended to be a complete coverage of all areas of proteomics, but rather to serve as an overview in order to provide an understanding of the work detailed in the following chapters.

I.1 Mass Spectrometry-Based Proteomics

I.1.1 Overview

Proteomics as a discipline can be defined as the identification and quantification of the complete set of proteins in a cell or tissue at a particular state. Although a number of alternative proteomics strategies such as protein array based methods have been developed, mass spectrometry (MS)-based proteomics has become the method of choice for large-scale studies. The applications of MS-based proteomics approaches have proved to be successful in molecular and cellular biology research including post-translational modification (PTM) identification and protein-protein interactions (Aebersold & Mann 2003). With recent improvements in instrumentation and methodology, proteomics has undergone tremendous advances over the past few years, enabling many powerful applications such as functional analysis of complex organisms (Schrimpf et al. 2009),

global analysis of PTM (Witze et al. 2007), large-scale reconstruction of protein interaction networks (Gstaiger & Aebersold 2009) and introduction of proteomics in clinical and translational research (Bousquet-Dubouch et al. 2011).

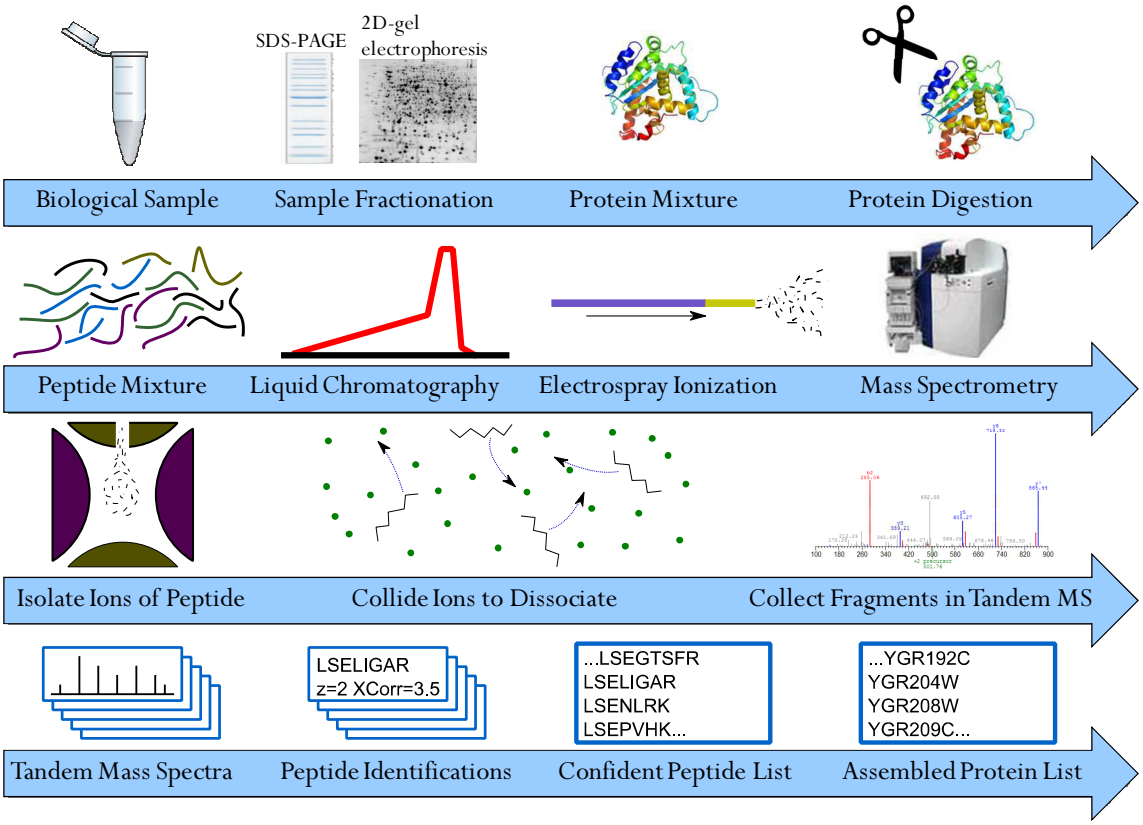

Figure 1. The typical MS-based proteomics workflow.

The typical workflow for a bottom-up MS-based proteomics experiment is illustrated in Figure 1. The first step is to reduce the complexity of a biological sample by one or several separation techniques such as SDS-PAGE and two-dimensional (2D) gel electrophoresis. Large proteins are then digested to peptides using site-specific proteases. Next, peptide mixtures are separated by liquid chromatography and ionized in a mass spectrometer. Precursor ions with particular mass-to-charge (*m/z*) values are selected and

collided with nonreactive gas to generate fragment ions. The corresponding *m/z* values and peak intensities of fragment ions are recorded in tandem mass spectra, which are interpreted to peptides by computational tools. Finally, the identified peptides are assembled into a list of proteins that are most likely present in the sample.

I.1.2 Sample Preparation and Separation

In proteomics studies, complex biological samples that contain a large number of proteins are often separated to simple mixtures prior to MS analysis. Various separation techniques can be used for this purpose. A widely used approach is to separate protein mixtures by SDS-PAGE, and then cut the gel to fractions for MS analysis. Samples of high complexity are now often fractionated by 2D-gel electrophoresis (Kenrick & Margolis 1970), which separates proteins based on their isoelectric points and molecular weights. Each spot in the gel may represent one or several purified proteins that can be further analyzed by MS. Recently a gel-based peptide-level isoelectric focusing approach (Hörth et al. 2006) has been shown to provide complementary coverage to the conventional gel-based fractionation method and yield higher identification rates (Hubner et al. 2008).

A gel-free approach known as shotgun proteomics directly analyzes large mixtures of peptides by coupling the electrospray ionization (ESI) of mass spectrometer in-line with a liquid chromatography (LC) system. Peptides are separated in the chromatography system to reduce the complexity. Two major types of LC systems are reverse phase high pressure liquid chromatography (RP-HPLC) that separates molecules by hydrophobicity and ion exchange chromatography that separates molecules by their

charges. High complexity samples can be separated using the multidimensional protein identification technology (MudPIT) (Washburn et al. 2001), which consists of a two dimensional chromatography. The first dimension is usually a strong cation exchange (SCX) column with high loading capacity. Eluted samples are subsequently separated by a reverse phase chromatography.

An alternative approach is the use of affinity chromatography to selectively enrich certain types of peptides or proteins. Affinity chromatography is often used to enrich post-translational modified peptides or proteins to make them more measurable by mass spectrometers. For example, the immobilized metal ion affinity chromatography (IMAC) can be used to enrich phosphopeptides (Thingholm et al. 2009), and blended antibody columns can be used to deplete plasma samples before MS analysis (Dayarathna et al. 2008, Pernemalm et al. 2009), which is a very effective way to reduce the sample dynamic range.

I.1.3 Protein Digestion

Proteins are usually cleaved to peptides by high specificity proteases prior to MS analysis. Trypsin is by far the most commonly used protease that cleaves peptides at the C-terminal side of arginine and lysine. Most proteins have tryptic cleavage sites that produce peptides with proper length for MS analysis. The cleavage generates “tryptic peptides” if both ends of peptide sequences conform to the trypsin cleavage rules. Specific cleavage on only one end of peptide sequences produces “semi-tryptic peptides”. Sometimes the “missed cleavages” may occur if resulting peptides contain internal trypsin cleavage sites.

The trypsin cleavage leaves a basic residue at the C-terminus which allows for a positive charge in acidic solution, producing charged peptides for MS analysis. Alternative site-specific protease such as chymotrypsin, GluC, LysC and AspN may also used in proteomics experiments, mainly for the increase of sequence coverage to distinguish homologous proteins or map PTM.

I.1.4 Mass Spectrometry Instruments

A mass spectrometer consists of three components: an ionization source, a mass analyzer and a detector. Peptides eluted from the LC system are transformed to gas phase charged ions, and then separated by mass analyzers with respect to their *m/z* values. Finally, the detector records the ions passing through mass analyzers, and reports them as mass spectra with *m/z* values of detected ions on the horizontal axis and their intensities on the vertical axis.

The ionization source introduces analytes into the instrument by transforming peptides or proteins to charged gaseous ions. Two major types of ionization methods in proteomics studies are matrix-assisted laser desorption/ionization (MALDI) (Tanaka et al. 1988) and ESI (Fenn et al. 1989). MALDI method co-crystallizes analytes with a matrix and applies UV laser light to vaporize them to charged ions. ESI sprays analytes to small droplets under high voltage. These droplets are subsequently vaporized to charged ions. Typically ions generated from MALDI are singly charged and ESI produces both singly and multi-charged ions.

The mass analyzer separates the charged ions based on their *m/z* values. In a bottom-up LC-MS/MS experiment, tandem mass spectra (MS/MS) are achieved by

performing two mass analyses. The first MS analysis measures the *m/z* values of ions (precursor ions), and selects ions in a certain range to undergo fragmentation. The selection can be controlled by instrument software. An exclusion list that contains the *m/z* values of most recently fragmented precursor ions can be used to reduce sampling redundancy. The resulting ions (product ions or fragment ions) are separated in the second mass analysis to generate tandem mass spectra.

Common mass analyzers used in proteomics experiments include quadrupole, ion trap, time of flight (TOF), Fourier transform ion cyclotron resonance (FTICR) and orbitrap. Each instrument has its strengths and weaknesses with respect to the speed, mass accuracy and resolution. More detailed discussions of these instruments are available in recent reviews (Yates et al. 2009, Chalkley 2010).

The ion trap instrument is probably by far the most widely used mass spectrometer due to its robustness, high sensitivity and relatively low price. However, the mass accuracy of ion trap is relatively low. In addition, there is a trade-off between the depth of the trapping potential and the width of the *m/z* range. Hence, in order to still contain the precursor ions, the *m/z* range has to be compromised. Usually ions below 1/3 of the precursor ion *m/z* will not be scanned in MS/MS, which is known as “low mass cut-off” of ion trap. For example, a peptide with 10 amino acids may have a neutral mass as 1100 Da. Even it is doubly charged, the ions below 183 *m/z* may not be acquired. In contrast, the mass range of immonium ions of amino acids is from 30 to 159. Therefore, immonium ions are often not observed in ion trap.

A recent major breakthrough is the proliferation of the LTQ-Orbitrap mass spectrometer (Hu et al. 2005). This hybrid instrument combines the robustness and

sensitivity of ion trap instruments with very high resolution and mass accuracy capabilities. It also has a higher dynamic range than FTICR (Makarov et al. 2006). In addition, LTQ-Orbitrap instruments can be configured to preserve low mass ions that are not observed in ion traps (as discussed in next section). The fast sequencing speed, high mass accuracy and high dynamic range make it particularly suitable for both qualitative and quantitative analysis of complex peptide mixtures (Olsen et al. 2009).

The mass accuracy and resolution of mass spectrometers have a substantial effect on the collected spectra. High mass accuracy also enables accurate determination of peptide ion charge state, thus greatly benefits the subsequent data analysis. It has been observed that data produced from high mass accuracy instruments can be better interpreted by bioinformatics tools (Zubarev & Mann 2007).

I.1.5 Peptide Fragmentation

Fragmentation Methods

In LC-MS/MS experiments, selected precursor ions are fragmented to product ions before detection. Figure 2 illustrates possible ions fragmented along the peptide backbone. The ion type depends on where peptide breaks and which side of the fragment receives the proton(s). If the charge is retained on the N–terminal side of the fragmented peptide, a, b or c ions are created, while x, y or z ions are generated if the charge is on the C-terminal side.

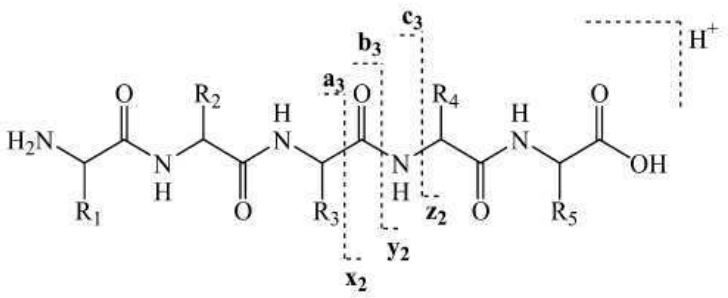

Figure 2. Theoretical fragmentation of a peptide. Adapted from Figure 2 in Wysocki et al. (2005).

Collision-Induced Dissociation (CID) is currently the most commonly used fragmentation method. Low-energy CID that is often used in quadrupoles and ion trap instruments mainly generates a, b, y ions and their neutral losses of water or ammonia. CID is a sensitive method and works well for low charged peptides (+2 or +3). However, labile modifications such as phosphorylation and glycosylation often lost during CID. In addition, it does not fragment long peptides well. These disadvantages can be solved by introducing Electron Transfer Dissociation (ETD) (Coon et al. 2005). ETD produces sequence-independent fragmentation and generates c and z ions. It particularly works well for long peptides, which can be generated by using other proteases instead of trypsin. Although ETD has lower sensitivity than CID, it preserves the labile modifications, making it a valuable method for phosphorylation and glycosylation studies.

Since CID works better for short peptides while ETD excels for long peptides, these two fragmentation methods therefore complement each other. A “decision-tree” model (Swaney et al. 2008) has been developed to assess peptide ions on-the-fly and determine which fragmentation method should be applied to these ions. This approach produced almost 40% more peptide identifications compared to CID alone.

Another fragmentation method is the Higher-energy Collision Dissociation (HCD) that is available in LTQ-Orbitrap instruments. It is particularly useful to pinpoint modifications such as phosphorylation because the immonium ions generated from HCD fragmentation will be preserved in mass spectra (Olsen et al. 2007). In addition, other low mass ions missing in ion trap instruments can be detected in LTQ-Orbitrap via HCD fragmentation, producing more abundant peaks in mass spectra. The high mass accuracy and abundant ions in the HCD spectra may greatly facilitate the downstream peptide identification (Bereman et al. 2011).

CID fragmentation is well supported by almost all peptide identification tools, while software for the analysis of ETD fragmentation data is currently less developed, and not all identification tools are now fully optimized to handle ETD data. Recent efforts have been made to develop new scoring methods specifically for the analysis of ETD spectra (Sadygov et al. 2009, Sun et al. 2010). A study also showed that an optimized scoring algorithm for ETD data can dramatically increase spectral identifications (Baker et al. 2010).

Understanding Fragmentation Pathway

The gas-phase peptide fragmentation process has not yet been fully understood. A number of studies have been conducted to investigate the fragmentation pathway (Wysocki et al. 2000, Zhang 2004, 2005, Klammer et al. 2008). The “Mobile Proton Model” (Wysocki et al. 2000) describes the fragmentation pathway under low-energy collision. In an ion trap instrument, for example, protonated precursor ions are trapped and undergo precursor ion selection, fragmentation, and fragment ion detection in the

same space. During CID, an ion trap applies a “tickle” RF voltage to induce peptide fragmentation. Under this voltage, precursor ions are excited to a higher internal energy level by collisions with nonreactive gas, making the charged proton migrating to energetically less favored protonation sites, such as peptide backbone. With a proton at the carbonyl oxygen of an amide bond, the preceding carbonyl can serve as a nucleophile to attack this carbonyl oxygen, forming an intermediate ring structure that subsequently breaks to dissociate the peptide bond (see Figure 3). The N-terminal fragment forms a b ion and C-terminal fragment becomes a y ion. This “charge directed” fragmentation occurs simultaneously in many molecules of the same peptide, resulting in different b and y ions that can be detected in MS/MS scan.

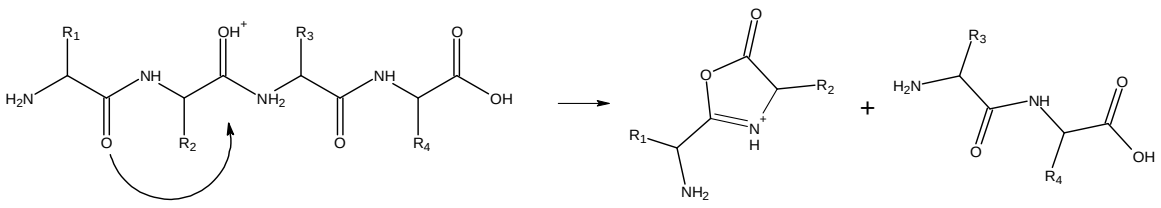

Figure 3. Mobile proton model for peptide fragmentation.

While “charge directed” peptide fragmentation is dominated in CID, peptide may dissociate in “charge remote” way that does not require the migration of a proton to peptide bond. The “pathways in competition” model (Paizs & Suhai 2005) explains several alternative fragmentations. For example, the side chains of aspartic acid, glutamic acid, asparagine, glutamine, histidine, lysine and arginine can attack their C-terminal carbonyls to break the peptide bonds and form b and y ions. Loss of water may occur in the C-terminal COOH group, N-terminal glutamic acid or serine/threonine containing

peptides. Loss of ammonia may occur from the side chains of asparagine, glutamine, lysine and arginine residues when the side chains are protonated. Peptides with labile PTMs often lose the modification groups because this process requires lower energy than breaking peptide bond. In low-energy CID, moving of proton(s), nucleophilic attack, breaking and forming chemical bonds are the principle chemical reactions that produce fragment ions.

Understanding the rules underlying the gas-phase peptide dissociation is important for the development of software tools. Current peptide identification tools often either implement a simple prediction model or totally ignore the intensities of product ions in their scoring schemes. Improving the prediction of product ion intensities increases the discrimination power of scoring systems for peptide identification (Havilio et al. 2003, Elias et al. 2004, Frank 2009a, b).

I.2 Proteomics Data Analysis

I.2.1 Overview

Automated bioinformatics tools play essential roles in proteomics data analysis (Domon 2006, Nesvizhskii et al. 2007). Frequently hundreds of thousands of tandem mass spectra are generated in a single proteomics experiment. The vast numbers of spectra place a heavy burden on data analysis, requiring an automated high throughput way for spectral interpretation.

Figure 4 summarizes the typical proteomics data analysis workflow. It starts with assigning peptide sequences to experimental spectra, which can be done with different strategies discussed in next section. Next, peptide identifications are validated to estimate

the confidence of the assignments, and high confident identifications are used to infer proteins. In many studies such as PTM analyses, advances searches may be conducted to interpret spectra that are evaded in the first round of analysis.

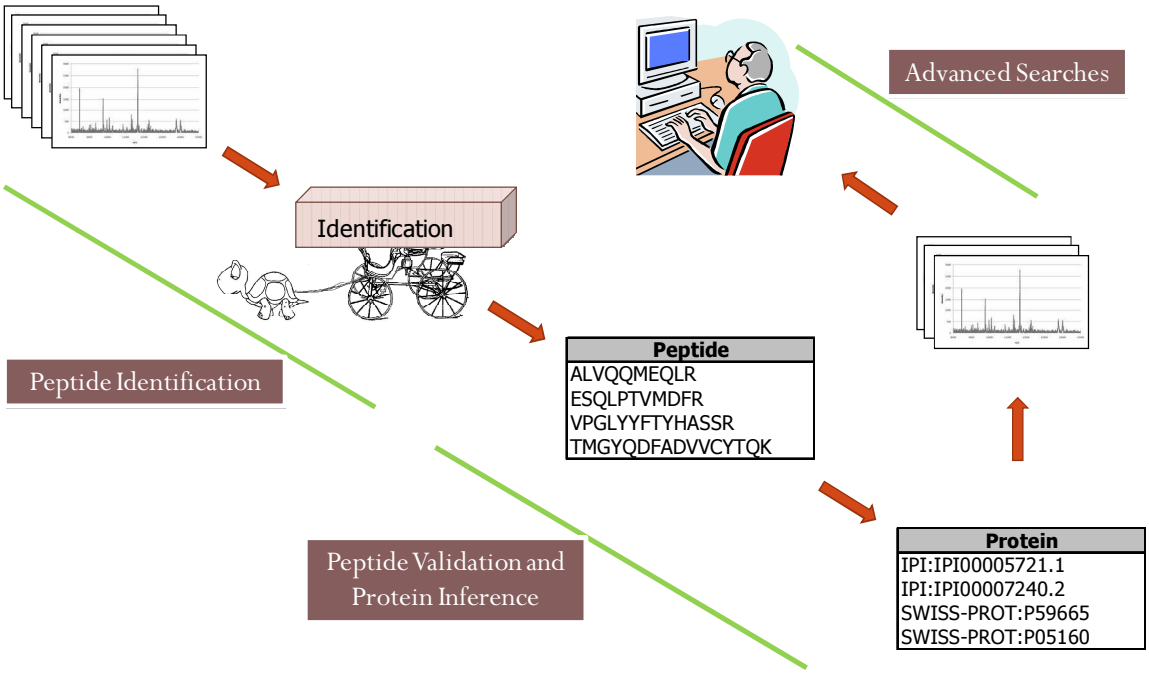

Figure 4. The typical MS-based proteomics data analysis workflow.

Bioinformatics tools have been used for MS-based proteomics data analysis since 1990s. During the past few years, many scoring algorithms have been developed to take advantage of improvements in MS instrumentation and fragmentation technologies. A partial list of these tools is summarized in Table 1.

Program	Web site	Reference
Database search tools		
Sequest	thermo.com	(Eng et al. 1994)
Mascot	matrixscience.com	(Perkins et al. 1999)
ProteinProspector	prospector.ucsf.edu	(Clauser et al. 1999)
SpectrumMill	www.chem.agilent.com	
Phoenyx	www.genebio.com/products/phenyx	(Colinge et al. 2003)
X!Tandem	www.thegpm.org	(Craig & Beavis 2004)
OMSSA	pubchem.ncbi.nlm.nih.gov/omssa	(Geer et al. 2004)
VEMS 3.0	yass.sdu.dk	(Matthiesen et al. 2005)
MyriMatch	fenchurch.mc.vanderbilt.edu/software.php	(Tabb et al. 2007)
ProteinPilot	www.absciex.com	
pFind 2.0	pfind.ict.ac.cn	(Wang et al. 2007)
Mass Matrix	www.massmatrix.net/mm-cgi/home.py	(Xu & Freitas 2008)
Andromeda	www.biochem.mpg.de/en/rd/maxquant	(Cox et al. 2011)
MassWiz	sourceforge.net/projects/masswiz	(Yadav et al. 2011)
***De novo* sequencing tools**		
Lutefisk	www.hairyfatguy.com/Lutefisk	(Johnson & Taylor 2002)
PEAKS	www.bioinformaticssolutions.com	(Ma et al. 2003)
Sequit	www.sequit.org	
PepNovo	proteomics.ucsd.edu/Software/PepNovo.html	(Frank & Pevzner 2005)
pNovo		(Chi et al. 2010)
Vonode	compbio.ornl.gov/Vonode	(Pan et al. 2010)
LysNDeNovo	gforge.nbic.nl/projects/lysndenovo	(van Breukelen et al. 2010)
Sequence tagging-based database search tools		
Popitam	www.expasy.org/tools/popitam	(Hernandez et al. 2003)
InsPecT	proteomics.ucsd.edu/Software/Inspect.html	(Tanner et al. 2005)
ByOnic	www.parc.com/work/focus-area/mass-spectra-analysis	(Bern et al. 2007)
MODi	http://modi.uos.ac.kr/modi	(Na et al. 2008)
TagRecon	fenchurch.mc.vanderbilt.edu/software.php	(Dasari et al. 2010)
Spectral library search tools		
X!Hunter	h201.thegpm.org/tandem/thegpm_hunter.html	(Craig et al. 2006)
Biblispec	proteome.gs.washington.edu/software/bibliospec/documentation/index.html	(Frewen et al. 2006)
SpectraST	www.peptideatlas.org/spectrast	(Lam et al. 2007)
Pepitome	fenchurch.mc.vanderbilt.edu/software.php	(Dasari et al. 2012)
Peptide validation and protein inference tools		
PeptideProphet	www.proteomecenter.org/software.php	(Keller et al. 2002)
ProteinProphet	www.proteomecenter.org/software.php	(Nesvizhskii et al. 2003)
MS-GF	proteomics.ucsd.edu/Software/MSGeneratingFunction.html	(Kim et al. 2008)
MaxQuant	www.biochem.mpg.de/en/rd/maxquant	(Cox & Mann 2008)
IDPicker	fenchurch.mc.vanderbilt.edu/software.php	(Ma et al. 2009)
Scaffold	www.proteomesoftware.com	(Searle 2010)
MassSieve	www.ncbi.nlm.nih.gov/staff/slottad/MassSieve	(Slotta et al. 2010)
PeptideClassifier	www.mop.unizh.ch/software.html	(Qeli & Ahrens 2010)

Table 1. Bioinformatics tools for MS-based proteomics data analysis.

I.2.2 Peptide Identification

The first step of data analysis is to assign peptide sequences to experimental spectra. As shown in Figure 5, the peptide identification strategies can be roughly summarized to four categories: database search, *de novo* sequencing, sequence tagging-based database search and spectral library search.

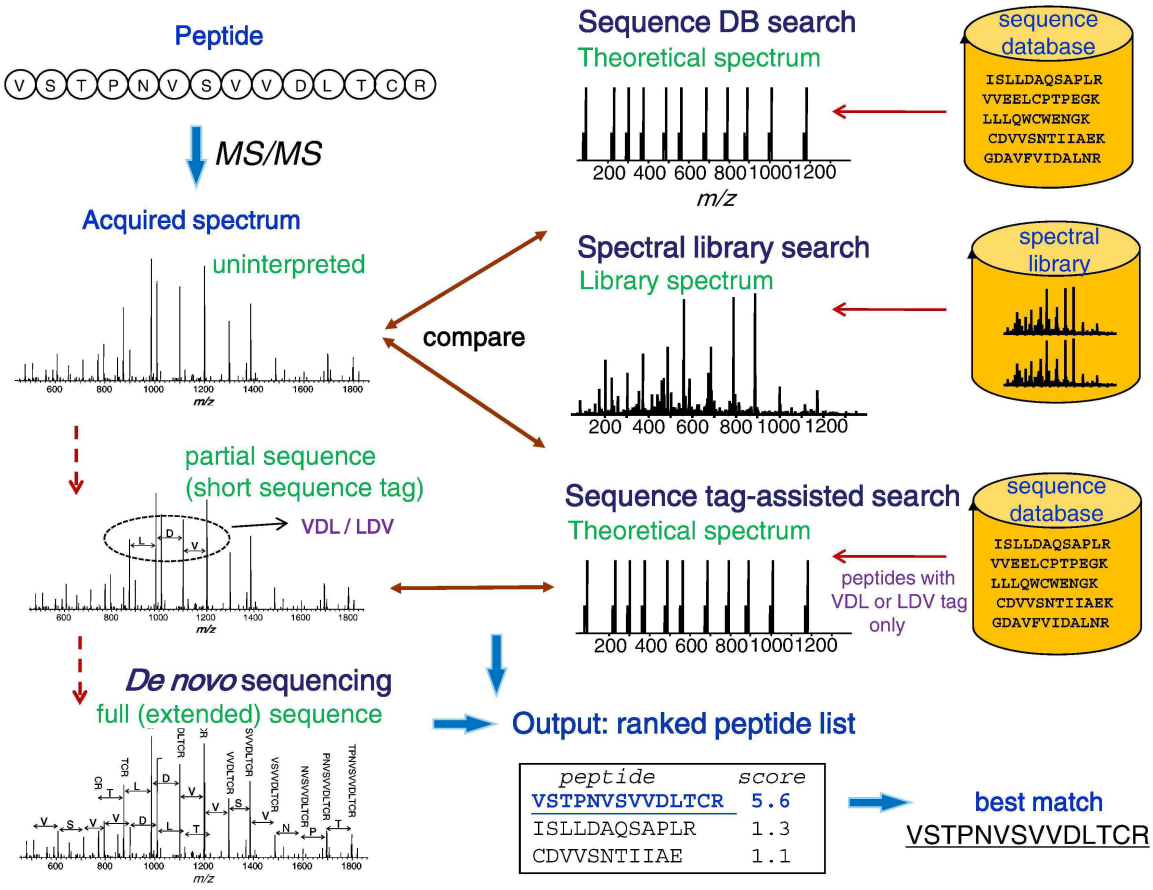

Figure 5. Four peptide identification strategies. Adapted from Figure 2 in Nesvizhskii (2010).

Database Search

The most widely used approach for peptide identification is to conduct a database search using software tools such as Sequest (Eng et al. 1994), Mascot (Perkins et al.

1999), X!Tandem (Craig & Beavis 2004), OMSSA (Geer et al. 2004) and MyriMatch (Tabb et al. 2007). Figure 6 illustrates the database search strategy for peptide identification. To interpret spectra, database search tools first perform an *in-silico* digestion of a protein database to enumerate all candidate peptide sequences, where masses of these peptides are similar to those of observed precursor ions. A theoretical spectrum constructed for each candidate sequence is then compared to the observed spectrum, producing a matching score to describe how well a peptide interprets the spectrum.

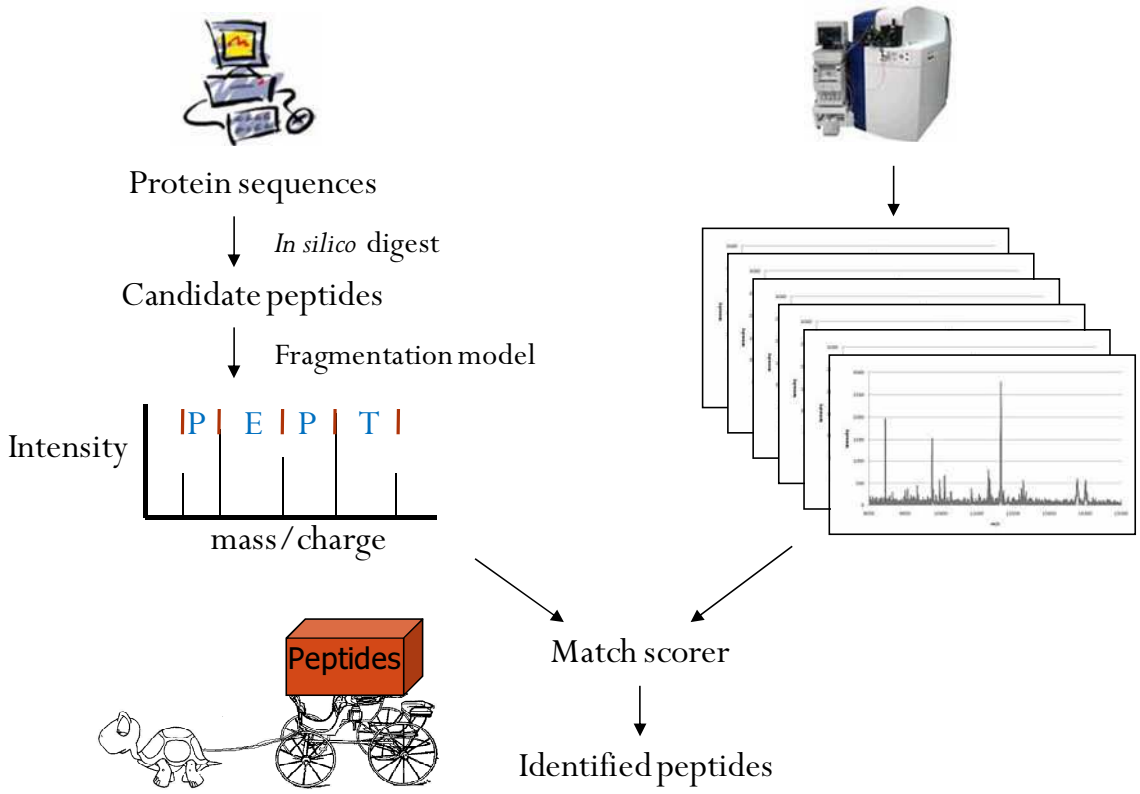

Figure 6. Peptide identification by the database search strategy.

The number of candidate peptides that are compared to a spectrum is affected by database search parameters, particularly precursor ion mass tolerance, enzyme digestion constraint and the number of allowed modifications (Nesvizhskii 2007). Although a large number of candidate peptides may be compared to a spectrum, database search tools usually only export the top few peptides ranked according to search scores. In most cases, only the top ranked peptide of each spectrum will be considered for the subsequent validation and protein inference.

A critical component in a database search program is the scoring function to measure the similarity between the experimental and theoretical spectra. A number of scoring schemes have been developed including the use of correlation functions (cross correlation in Sequest and dot product in X!Tandem) or probability-based models (Mascot and MyriMatch). Usually database search tools implement multiple scoring functions to evaluate the peptide-spectrum-matches (PSMs) in different aspects. These scores vary from arbitrary values such as XCorr in Sequest to statistical measures such as e-values in X!Tandem. Individual scores or the combination of multiple scores can be used for the subsequent peptide validation.

Database search parameters have a great impact on search results. First, the precursor mass tolerance determines which peptides will be compared to the experimental spectrum, i.e., only peptides with masses within the precursor mass tolerance will be scored. High mass accuracy instruments allow a very narrow mass window specified in database search compared to low mass accuracy data (e.g. 10 ppm for orbitrap data compared to 3 Da for LTQ). This leads to fewer possible candidate peptides that are compared to the observed spectrum, thus dramatically reduces searching

time and decreases the number of false matches. Second, enzyme digestion constraint also controls the number of candidate peptides to be compared. For example, a tryptic search produces less candidate peptides than an unconstrained or semi-tryptic search. As a result, it usually spends less time than non-tryptic searches. A tryptic search, however, eliminates the possibility to identify peptides that undergo unexpected cleavages. Meanwhile, other database search parameters such as the number of allowed modifications, deisotoping setting and the reference protein database can also affect the search results (Nesvizhskii 2010).

Although database search offers an automated high-throughput approach for peptide identification, they rely heavily on protein databases, in which some of the genome sequences and annotations may not be accurate. More importantly, mutations and modified peptides in biological samples are often ignored by existing database search methods. In addition, database search is a very time-consuming process because the large number of comparisons between observed spectra and their candidate peptides. These issues are addressed by the development of the ScanRanker tool described in Chapter III.

De Novo Sequencing

Unlike database search that requires a reference protein database for peptide identification, *de novo* sequencing infers peptide sequences directly from experimental spectra. The inferred peptides can be mapped to proteins by downstream tools such as MS-BLAST (Shevchenko et al. 2001). This is particularly useful when the organisms of interest have unsequenced or partially sequenced genomes. However, since this approach requires high spectral quality for accurate interpretation, and is very computationally

intensive, it has not yet been used for large-scale proteomics data analysis. The ScanRanker tool described in Chapter III helps to alleviate this problem.

As summarized in Table 1, several *de novo* sequencing tools have been described. Early tools such as PepNovo (Frank & Pevzner 2005) and PEAKS (Ma et al. 2003) were developed for low resolution data under CID fragmentation. Recent efforts have been made to develop new *de novo* sequencing algorithms for high mass accuracy data (Frank et al. 2007, Pan et al. 2010) or data collected under HCD (Chi et al. 2010) and ETD (van Breukelen et al. 2010) fragmentation. These researches demonstrated that *de novo* sequencing can be greatly improved by the use of high mass accuracy instruments and advanced fragmentation methods.

Sequence Tagging-Based Database Search

Sequence tagging-based database search combines *de novo* sequencing and database search strategies. It first infers short peptide sequences ("tags") from spectra. These tags are then used to match candidate peptides via database search. A tag comprises three parts in mass-sequence-mass format: the mass flanking the N-terminal of the partial sequence, the partial sequence, and the mass flanking the C-terminal of the partial sequence. A candidate peptide is selected to score against the spectrum if both the partial sequence and flanking masses in the observed spectrum match to the peptide. Compared to traditional database search methods that use precursor masses to select candidate peptides, sequence tagging employs tags as the text-based filter, which improves specificity and reduces the number of candidate sequences by a few orders of magnitude.

Sequence tagging-based approach is particularly useful for the identification of mutations or post-translationally modified peptides (Mann & Wilm 1994, Nesvizhskii 2010). Bioinformatics tools such as InsPecT (Tanner et al. 2005), MODi (Na et al. 2008) and TagRecon (Dasari et al. 2010) are examples that employ sequence tagging to enable modification searches. These programs treat the mass shifts between experimental spectra and candidate peptides as potential modifications, and place the mass shifts on amino acids that best explain the spectra. Both *de novo* sequencing and sequence tagging-based database search benefit from the high mass accuracy of modern mass spectrometers. In Chapter III, I will discuss the use of sequence-tagging approach for spectral quality assessment.

Spectral Library Search

Spectral library search is a fast and sensitive approach for peptide identification compared to a conventional database search. Rather than matching observed spectra to computationally modeled theoretical spectra, MS/MS scans can be interpreted by matching against a spectral library, which is a large collection of observed spectra that are confidently identified in previous experiments. Bioinformatics tools such as SpectraST (Lam et al. 2007), Bibliospec (Frewen et al. 2006), X!Hunter (Craig et al. 2006) and Pepitome (Dasari et al. 2012) were developed for spectral library searching. The National Institute of Standards and Technology (NIST) made several spectral libraries publically available for multiple species (http://peptide.nist.gov).

Spectral library search is very computationally efficient. The accuracy of this method is considered to be higher than conventional database search. It is particular

useful for fast identification of well-studied samples. For example, bovine serum albumin (BSA) samples are routinely analyzed for instrumentation quality control (QC). Spectral library search is an ideal method for quick identification of these QC samples. A disadvantage of spectral library search is that only peptides that are previously identified can be assigned to newly observed spectra, and its performance is largely affected by the completeness and accuracy of assembled spectral libraries. In addition, a spectral library constructed for a particular type of mass spectrometer may not be applicable to data collected on other types of instruments due to the different gas-phase fragmentation principles.

I.2.3 Peptide Validation

Overview

Peptide identification tools evaluate all possible candidate peptides for each input spectrum, and usually only the best-scoring sequence is used to interpret the spectrum. However, not all PSMs are correct assignments. In contrast, sometimes the majority of best-scoring peptides assigned by database search tools are incorrect PSMs (Domon 2006, Nesvizhskii et al. 2007). The reasons for the high failure rate include:

(1) Sequence not in database. Peptides with mutations and unexpected modifications will not be identified. Their spectra may be assigned incorrect best-scoring peptides.

(2) Contaminant spectra. Database search only identifies spectra derived from peptides, while chemical contaminants that are introduced to MS analysis during sample preparation are assigned wrong peptide sequences.

(3) Low quality spectra. Poorly fragmented peptides often produce low quality spectra that have either high signal-to-noise (S/N) ratio or less peaks to match peptide sequences, thus may be assigned incorrect peptides in database search.

(4) Insufficient scoring scheme. Database search engines often apply a simplified fragmentation model to predict the theoretical spectrum, while in reality peptide fragmentation depends on many factors such as amino acid composition and location, and produces more complicated spectra.

(5) Chimera spectra. Multiple peptides with the same *m/z* value may be concurrently isolated at the same time, thus produce a chimera spectrum with fragment ions from all these peptides. Database search tools may assign one of the correct peptides or a wrong sequence to a chimera spectrum.

(6) Incorrect precursor charge state or mass. The precursor ion mass of a spectrum can be measured inaccurately, and wrong candidate peptides may be selected to match the spectrum. Meanwhile, peptide charge state can be incorrectly determined, especially for low resolution instruments such as LTQ.

(7) Inappropriate search parameters. A wide precursor mass tolerance introduces more candidate peptides for comparison, thus has a potential to produce more incorrect PSMs. A narrow precursor mass tolerance has the risk to exclude correct peptides for comparison. A tryptic search will not identify peptides with unexpected cleavage, resulting in incorrect peptides assigned to these spectra.

Figure 7 illustrates the score distribution of correct and incorrect PSMs, which may overlap significantly depending on how well they can be discriminated by database

search scores. Incorrect PSMs may score higher than some correct PSMs due to spurious matches, homologous peptide sequences, or because spectra for these correct PSMs are relatively low quality spectra. It is desired that database search programs achieve a high discrimination between correct and incorrect PSMs in peptide identification. Improving database search scoring schemes and developing advanced peptide validation methods may both reduce the overlap region of correct and incorrect PSMs, and subsequently reduce false peptide sequences for protein inference.

Some correct PSMs may be excluded for subsequent analysis because they fail to pass the confidence threshold (see Figure 7, region A'). Meanwhile, some spectra are assigned incorrect peptides because these peptides are scored better than correct ones due to many possible reasons described above. These issues can be alleviated by the introduction of the IDBoost tool described in Chapter II.

Since peptides with unexpected modifications and mutations will not be identified in database search, these spectra will generate incorrect PSMs. Advanced identification methods such as sequence tagging-based modification search or *de novo* sequencing helps to interpret these spectra, while how to find these spectra remains an issue. In Chapter III, I will demonstrate the use of the ScanRanker tool to solve this problem.

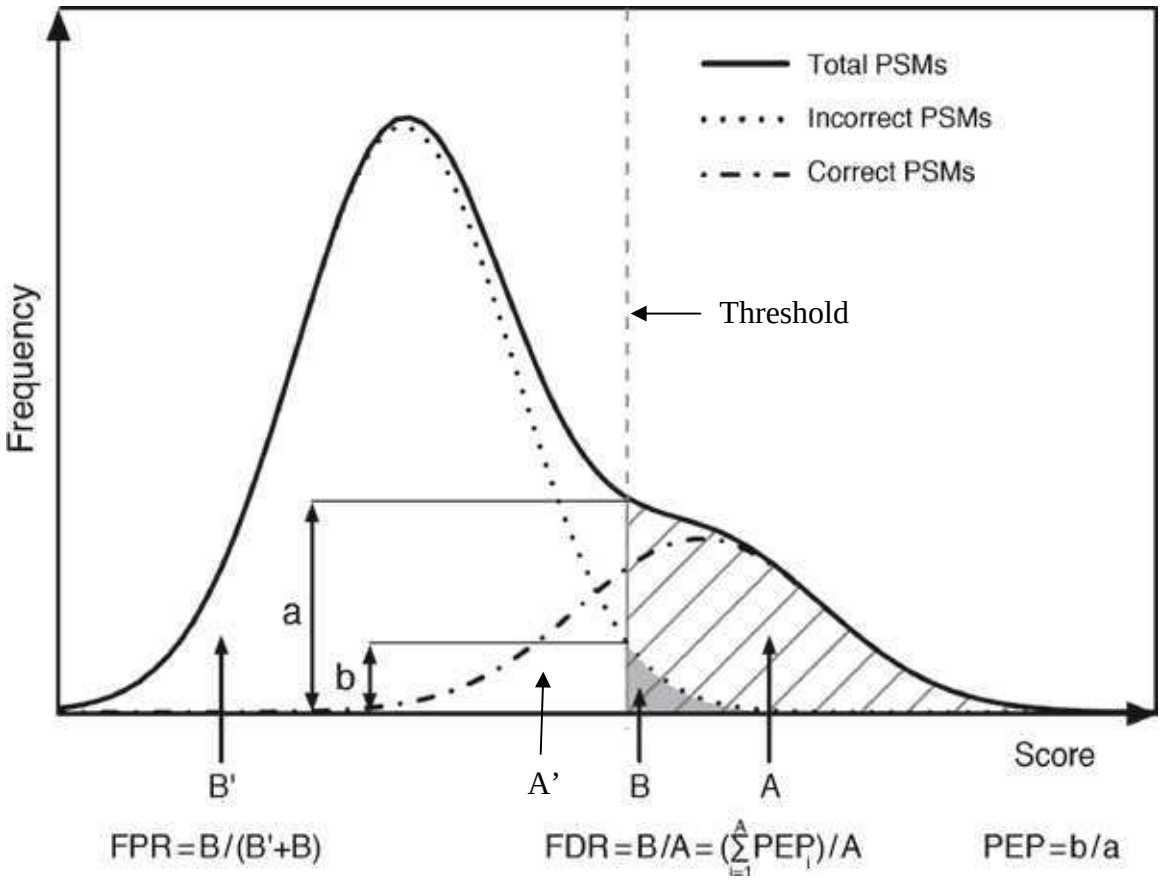

Figure 7. Score distribution for correct and incorrect PSMs. Adapted from Figure 1 in Brosch & Choudhary (2010). Shaded area (a) represents all accepted PSMs (both correct and incorrect PSMs) above a threshold, and solid grey filled area (b) represents incorrect PSMs passing the threshold that are falsely accepted. A' together with A sum up all correct PSMs, and A' represents correct PSMs that fail to pass the threshold. B and B' sum up all incorrect PSMs, and B represents incorrect PSMs that are wrongly selected within a given threshold. The false positive rate (FPR), false discovery rate (FDR) and posterior error probability (PEP) can be calculated as shown in the figure.

Peptide Validation Strategies

Because a large proportion of MS/MS spectra cannot be matched successfully to peptide sequences, raw identifications must be filtered to retain the most accurate PSMs for protein inference, i.e., a threshold need be determined to generate a list of high confident identifications. The selected threshold should yield a good tradeoff between sensitivity and error rate. A high score threshold reduces the number of false matches but

also decreases sensitivity, yielding less number of correct PSMs for protein inference. In contrast, a low score threshold allows more PSMs to be selected at the cost of a higher error rate.

Early on proteomics researchers often applied an *ad hoc* cutoff value of database search scores to generate a list of confident PSMs. For example, use XCorr > 2.5 for Sequest search and IonScore > 45 for Mascot search. This approach, however, has many disadvantages. First, the score distributions generated by a database search tool vary with respect to the instruments, sample complexities, data quality and the protein database searched. Therefore, there is no single score threshold can be applied to all datasets. Second, even though a single score threshold can be applied to data from different experiments, the error rates are still remaining unknown, making it difficult to compare data between experiments. Third, applying an *ad hoc* cutoff makes it impossible to compare search results from different search algorithms and instruments, and often has poor tradeoff between sensitivity and specificity.

To solve these issues, modern proteomics has moved away from the *ad hoc* score cutoff toward probabilistic approaches. Translating the database search scores to statistics provides interpretable probability scores. Multiple search scores, database features and experimental conditions all can be taken into account in statistical models.

Several methods have been developed to convert arbitrary search scores of raw identifications into statistical measures. As shown in Figure 7, three commonly used statistical measures are p-value, false discovery rate (FDR) and posterior error probability (PEP).

Database search scores can be converted to p-values to measure the confidence for peptides scored to a single spectrum. In order to interpret a spectrum, database search engines enumerate all candidate peptides, and each of them is scored against the spectrum. This produces a large number of scores that can be used to estimate the null distribution for p-value inference. The score of the best matched peptide is then converted to a p-value based on the null distribution. Both parametric distribution (Sadygov & Yates 2003, Geer et al. 2004) and empirically fitted distribution (Fenyö & Beavis 2003) have been developed to derive p-values. A p-value can be interpreted as the probability to observe a match with an equal or higher score by random chance. Therefore, the further a score is away from the center of the null distribution, the higher the statistical significance it represents.

A disadvantage of the p-value approach is that it is affected by the number of PSMs compared to a spectrum. Large number of comparisons may yield smaller p-values by random chance alone, which requires a multiple testing correction to adjust p-values. However, classical methods such as “Bonferroni correction” were not designed for large size of datasets, and often lead to overly conservative results.

An alternative statistical measure that works well for large-scale data is FDR, which estimates the global error rate for a set of PSMs. In proteomics, for example, if 100 PSMs were scored above a threshold and 5 of them were found to be incorrect matches, then the expected FDR will be 5% for this analysis. A common way to estimate incorrect matches among a collection of PSMs is to conduct database searches through the target-decoy strategy (Elias & Gygi 2007). This approach searches MS/MS scans against a target protein database appended with decoy proteins, which can be reversed (Moore et al.

2002), randomized (Colinge et al. 2003) or shuffled (Klammer & MacCoss 2006) sequences. It assumes that false identifications follow the same distribution as matches to decoy sequences. To compute FDR, all PSMs from a database search are ordered by a matching score or a combination of multiple matching scores. A q-value is then calculated for each PSM as the minimal FDR threshold at which a PSM is accepted. PSMs passing a FDR threshold are then considered valid identifications for protein inference.

FDR-based peptide validation has become the method of choice for large-scale proteomics studies, and many bioinformatics tools have implemented this approach. In my Master's thesis, I presented IDPicker 2.0 that combines multiple search scores and applies additional filters to improve FDR-based peptide validation (Ma et al. 2009). Another tool, Percolator (Käll et al. 2007), employs a semi-supervised machine leaning method to discriminate between correct and incorrect PSMs based on target-decoy search results. It was originally designed to work with Sequest results and has been recently adapted to handle Mascot search results (Brosch et al. 2009).

Although q-values are associated with individual PSMs, FDR is a summary statistic for the entire collection of PSMs, and does not measure the confidence of individual PSMs. When the focus is to evaluate individual PSMs, PEP, also known as local FDR, can be estimated to represent the probability of a PSM being incorrect. For example, a PSM with a PEP value of 0.01 means there is 1% chance that this PSM is an incorrect assignment. One way to compute PEP is to use a mixture model-based method as implemented in PeptideProphet (Keller et al. 2002). The PEP for each individual PSM can be used to filter low confident identifications. Moreover, the PEP and FDR method

can work together to make more accurate and robust estimation (Choi & Nesvizhskii 2008). In this case, the decoy sequences are used to estimate the distribution of incorrect PSMs, yielding a more accurate mixture model for PEP calculation.

I.2.4 Protein Inference

In most proteomics experiments, the ultimate goal of a study is to know what proteins are present in the analyzed sample. Therefore, high confident peptide sequences passing the validation step need to be mapped to their corresponding proteins, and the confidence at the level of proteins need to be re-assessed. This process, however, is not straightforward and faces many challenges.

First, peptides whose sequences are present in more than one protein may complicate the protein inference process. In this case, since a single peptide can be mapped to multiple proteins, it is difficult to know which protein(s) is present in the analyzed sample. As illustrated in Figure 8, for example, protein B and C will be indistinguishable because they both map to the same set of peptides. The shared peptides often result from homologous proteins, splicing variants or redundant entries in the protein database. This is particularly a serious problem for higher eukaryote organisms due to the high abundance of shared peptides (Nesvizhskii & Aebersold 2005). It is a general problem for shotgun proteomics experiments because the connectivity between peptides and proteins is lost during sample preparation and digestion. Separating proteins in a 2D gel before MS analysis helps to alleviate this problem, where additional information such as the molecular weights and isoelectric point can be used in determination of the protein identities (Görg et al. 2004).

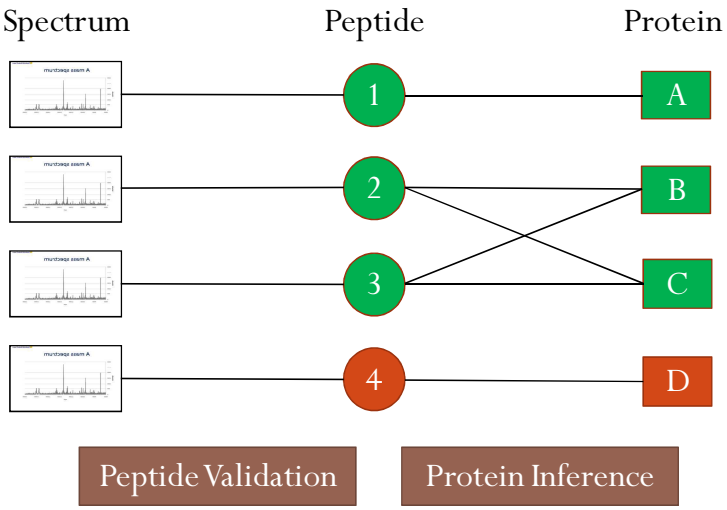

Figure 8. A simplified example of protein inference. Green and red colors represent correct and incorrect peptides/proteins, respectively. Peptide 2 and 3 are shared by the same set of proteins.

Second, incorrect PSMs may be accepted after peptide validation, yielding wrong peptides for protein inference (e.g. protein D in Figure 8). This is a more serious problem when searching spectra against a large protein database, where spurious peptides have a higher chance to be scored superior to correct ones. At the same time, many correct PSMs tend to map to a relatively small number of proteins that are dominant in the analyzed sample (Nesvizhskii et al. 2003). For example, a recent study showed that only ~5% of all collected MS/MS scans lead to the identification of unique peptides in large-scale studies (Swaney et al. 2010). As a result, almost every highly scored incorrect PSM may introduce one additional incorrect protein. Even with a careful control of FDR at the PSM level, these incorrect PSMs can produce a high FDR at the protein level. Requiring more than one distinct peptide per protein ("two peptide rule") helps to remove some

incorrect proteins, but this will reduce the sensitivity and excludes the identification of low abundance proteins supported by a single peptide.

A commonly used approach for protein inference and error rate estimation is to conduct database searches using a target-decoy strategy, and then apply various filters to assemble proteins to a desired protein-level FDR. Common filters include peptide-level FDR, minimal number of spectra per protein and minimal number of distinct peptides per protein. In this case, the protein-level FDR can be estimated according to the number of decoy proteins included in the final list. To achieve a lower protein-level FDR, one can either apply a more stringent peptide-level FDR, or require more than one spectrum or distinct peptides per protein. Both approaches lower the number of incorrect PSMs for protein inference, and thus reduce the error rate. To handle the problem of shared peptides that may produce many homologous proteins and isomers in final list, one can either report all proteins identified with at least one distinct peptide, or simply select a representative protein among homologs (States et al. 2006).

The parsimony principle for protein inference has been widely accepted in proteomics community. It is also required by several journals for publishing proteomics research results (Carr et al. 2004). The central concept, as exemplified by several computational tools (Nesvizhskii et al. 2003, Yang et al. 2004, Zhang et al. 2007, Ma et al. 2009), is to derive a minimal list of proteins that can account for all observed peptides.

A disadvantage of the protein-level FDR is that it is a global estimation of error rate for all accepted proteins. The confidence of individual proteins may be further estimated based on many metrics such as sequence coverage and the number of identified spectra for corresponding proteins. Statistical models have been developed to compute

probabilities for individual proteins, which estimate the likelihood that a protein is a true identification. For example, ProteinProphet (Nesvizhskii et al. 2003) reads the PSMs and their posterior probabilities generated from PeptideProphet to compute a cumulative score. That is, the probabilities of all PSMs mapped to a protein are combined together to yield the probability that the corresponding protein is present in the analyzed sample.

The initial PSM probabilities from PeptideProphet may be adjusted to take into account the number of peptides mapped to the same protein group (undistinguishable proteins). The adjustment produces improved protein probabilities that agree with the actual protein-level FDR. ProteinProphet retains proteins identified by a single peptide if that peptide is assigned a high posterior probability in PeptideProphet. These proteins could be excluded in FDR-based protein inference due to the use of "two peptide rule". Other statistical methods using hierarchical modeling (Shen et al. 2008) or incorporating gene models to protein inference (Gerster et al. 2010) were also reported.

Most bioinformatics tools separate peptide validation and protein inference to two steps as described above. A recent research treated protein inference as a single optimization problem, and proposed a machine learning method, Barista, to optimize these two steps in a single analysis (Spivak et al. 2011). The essential concept is that peptide validation and protein inference are cooperative such that one task benefits from the other during optimization, and thus should be exploited simultaneously. Barista reads target-decoy search results and develops a model that maximizes the number of target proteins. It incorporates a wide variety of evidence to directly control the relevant error rate, providing 18-34% more protein identifications than other approaches (Spivak et al. 2011).

I.3 Instrumentation Quality Control

No matter how advance the data processing algorithms could be, they all assume the spectra from mass spectrometers are collected under stable instrument performance. Therefore, quality control of instrumentation performance is critical for proteomics studies. Many studies are designed to be comparative in nature such as exploring protein expression differences between tumor and normal tissues. These studies assume the observed differences come from the proteome differences of analyzed samples rather than analytical system variability. Therefore, the mass spectrometer needs to be frequently checked during data collection to ensure stable analytical system performance. Even with high mass accuracy instruments, achieving truly high accuracy often requires fine instrument tuning, room temperature control and the use of internal or external calibration.

The most commonly used approach is to run simple samples such as BSA periodically, and count the number of confident identifications to measure instrument variability. This approach, however, does not reveal whether system performance is optimal or which components cause the large variation. NIST introduced the MSQC software (Rudnick et al. 2010) to compute diverse metrics from experimental LC-MS/MS data, enabling the QC evaluation of proteomics instrumentation. In practice, however, several aspects of the MSQC software prevent its use for routine instrument monitoring. This problem is further addressed in Chapter IV with the development of the QuaMeter tool.

I.4 Dissertation Outline

The objectives of my work are to develop novel algorithms and bioinformatics tools for MS-based proteomics data analysis. The following chapters present three tools that facilitate proteomics data processing. In each chapter a separate introduction is given to describe the background of the respective topic.

In Chapter II, I present the IDBoost tool to rescue correct spectral identifications and correct database search errors through spectral clustering. In Chapter III, I describe the ScanRanker tool that evaluates the quality of tandem mass spectra via the sequence tagging approach. In Chapter IV, I present the QuaMeter tool for MS instrumentation quality control. Each tool is evaluated with a variety of datasets and their applications are demonstrated.

CHAPTER II

IDBOOST: VALIDATION AND RESCUE OF TANDEM MASS SPECTRAL IDENTIFICATIONS VIA SPECTRAL CLUSTERING

II.1 Introduction

Despite recent improvements in analytical methods, usually only a small fraction of spectra can be identified in a typical shotgun proteomics experiment, implying the need for advanced methods to improve identification rate. This may be caused by many factors such as unexpected modifications, incomplete protein databases or low spectral quality. However, many spectra assigned correct peptides may fail to pass the FDR threshold (see Figure 7). For example, given a set of spectra assigned to the same peptide, it is common that only spectra assigned high database search scores are identified, while the others that fail to pass the threshold are discarded. These discarded spectra may be correct identifications because the matched peptide is identified by other spectra. Rescuing these spectral identifications provides more information for subsequent data analysis such as manual validation of phosphopeptides and spectral count-based protein quantification.

In addition to the low identification rate, two kinds of errors are often included in database search results. First, wrong peptides may spuriously score higher than correct sequences. This introduces false proteins and reduces the spectral count of correctly identified proteins, leading to inaccurate estimations in spectral count-based protein quantification. Second, if multiple modification sites are present in a peptide, one with a

misplaced modification site may score better than a correct one due to low spectral quality or insufficiency of scoring algorithms. The ambiguous modification locations are detrimental to experiments to localize modifications, such as phosphorylation studies. For both kinds of errors, the correct sequences frequently score very similarly to the erroneous top-ranked matches. Since many database search engines generate several PSMs per spectrum, it is very likely that the correct sequences are stored in the search output, but are invisible in subsequent analysis because they are not top-ranked hits. They can, however, be rescued by examining search results and re-ranking PSMs for each spectrum.

Several efforts have been made to correct these errors. For example, Percolator provided a re-ranking function to correct spurious random matches via a machine learning approach for Sequest or Mascot, deciding which PSM was ranked highest for a spectrum by search scores and peptide properties. Ascore (Beausoleil et al. 2006) presented a probability-based score to correct phosphorylation site localization, but it required the presence of site-determining ions exclusive to specific site locations. These methods correct errors based on search results from either a single file or a single spectrum. In fact, shotgun proteomics experiments are often designed to include multiple replicate LC-MS/MS runs, and many identified peptides are associated with more than one spectrum. For example, in a recent study only ~5% of all collected MS/MS scans lead to the identification of unique peptides (Swaney et al. 2010).

Here I seek to correct these errors in a single analysis by incorporating search results across multiple runs. I hypothesized that spectra derived from the same peptide should share high similarity in fragment ion patterns. Given a set of similar spectra, a

secondary PSM (ranked below the first position for a spectrum) may represent the correct interpretation if similar spectra also are matched to this peptide. Likewise, the modification site localization errors may be corrected by taking into account site assignments of similar spectra. This approach rescues correct secondary PSMs based on existing search results with no requirement for running additional database searches. For the best applicability, the approach must function with a variety of search engines and use more informative tandem mass spectra to guide interpretation of poorer quality scans.

In this work, I seek to rescue spectra that are supported by other confident PSMs passing the FDR threshold. However, simply adding all spectra assigned to these peptides back to the analysis is not appropriate, because some of them may be unreliable spurious matches. In addition, if multiple PSMs per spectrum are considered, more than one peptide could be identified and it is not clear which PSM should be rescued. Here I present IDBoost, a software tool to rescue spectral identifications and correct database search errors via spectral clustering. I demonstrate the use of IDBoost in phosphorylation studies to rescue phosphopeptide identifications and to resolve phosphosite localization ambiguity. I show that IDBoost helps recognize differentially expressed proteins in comparative analysis. I also evaluate IDBoost using a variety of datasets representing various instrument platforms and sample complexities.

II.2 Algorithm

II.2.1 Overview

The goal of this work is to rescue PSMs and to correct database search errors by incorporating identification evidence from similar spectra. In brief, IDBoost first groups

similar spectra into clusters and then examines all pairs of spectrum-peptide matches in a cluster. A PSM will be rescued if a similar spectrum matched to the same peptide is a valid identification. Multiple PSMs per spectrum, e.g., the top 5 ranked PSMs for a spectrum, can be included in this process, enabling re-ranking of PSMs to correct spurious matches or modification localization errors. Only one PSM per spectrum is allowed to be rescued. A "Bayesian average" rating method prioritizes peptides for rescue. IDBoost is written in C#/.NET and implemented in IDPicker (Zhang et al. 2007, Ma et al. 2009), which is available for download from http://fenchurch.mc.vanderbilt.edu.

II.2.2 Spectral Clustering

Tandem mass spectra are clustered based on the similarity between each pair of spectra. Rather than process all spectra, only spectra matching to a confidently identified peptide within the top N ranked PSMs are selected for clustering (N is a user configurable parameter). Next, selected spectra are sorted by their precursor *m/z* values and are compared for similarity to any others within a user-specified *m/z* tolerance. The similarity between each pair is computed by a normalized dot product, which has previously been found to work well for spectral clustering (Tabb et al. 2003, 2005, Beer et al. 2004, Frank et al. 2008). To reduce the effect of low intensity peaks, only the top 100 most intense peaks of each spectrum are retained for similarity comparison. Peak intensities are square rooted to emphasize smaller peaks (Tabb et al. 2003). A single-linkage clustering approach (Beer et al. 2004, Frank et al. 2008) is applied to group spectra. i.e., if spectrum A is similar to B, and B is similar to C, then all three spectra will form one cluster. The default similarity threshold is 0.6, and is user configurable. The

method is similar to the Pep-Miner algorithm (Beer et al. 2004), which has been proved to be an effective clustering approach in prior work (Beer et al. 2004, Frank et al. 2008).

II.2.3 Rescue of Spectral Identifications

Once similar spectra are grouped together, an unidentified PSM may be rescued by taking identification evidence from other spectra into account. As illustrated in Figure 9, for example, all spectra in a cluster are first mapped to peptides in a bipartite graph. Multiple PSMs per spectrum can be included to enable the rescue of secondary PSMs. Next, a "Bayesian average" rating method (described below) is applied to prioritize peptide sequences that will be processed. This is a necessary step because only one PSM per spectrum is allowed to be rescued, while one spectrum may be mapped to multiple peptides (the top N peptides assigned to this spectrum). Peptides sharing the same sequence but different PTM locations are treated as distinct peptides.

To rescue unidentified PSMs, IDBoost sifts through prioritized peptides and their linked spectra. An unidentified PSM will be rescued if a similar spectrum matched to the same peptide is a valid identification. A rescued spectrum then will be excluded from further analysis to ensure only one PSM per spectrum rescued. In Figure 9C, for example, the best scored Pep2 linked to three spectra. Since the PSM of Scan4-Pep2 is a valid identification, both Scan2 and Scan5 will be rescued. Next, Scan3 will be rescued to Pep1 because the same peptide is supported by an identified spectrum (Scan1). However, since only one PSM per spectrum is allowed to be rescued, Scan2 will not be assigned to Pep1 because it has already been processed and rescued to Pep2. Since no spectra linked to Pep3 are identified, all PSMs mapped to Pep3 remain untouched.

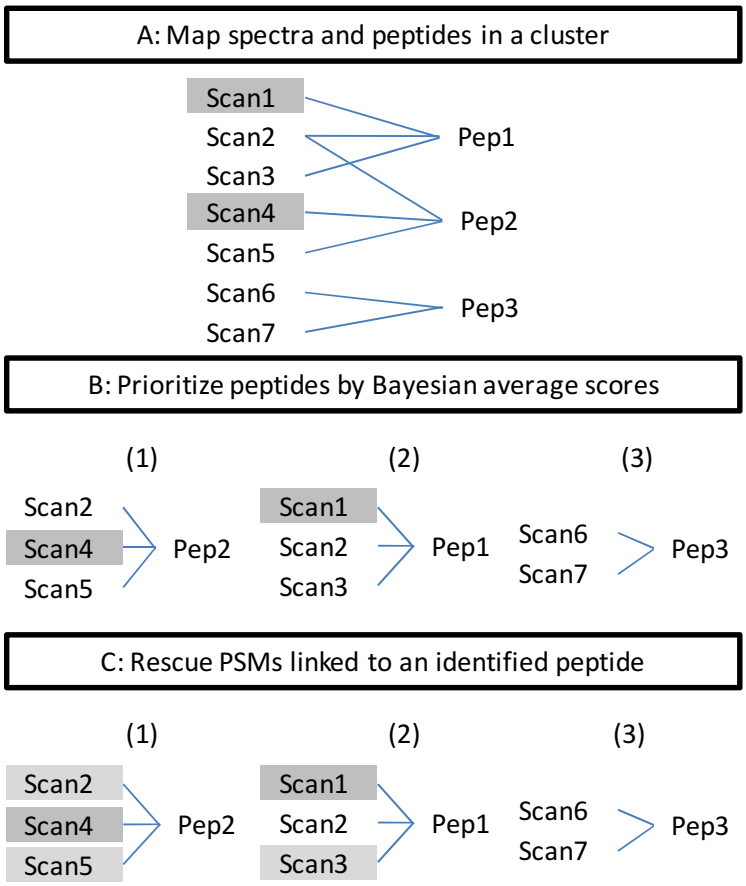

Figure 9. A diagram of rescuing unidentified spectra in a cluster. (A) A bipartite graph shows the PSM mapping between seven spectra and three peptides in a spectral cluster. Each link represents a peptide-spectrum match. For simplicity, all spectra are linked to only one peptide except Scan2, which is mapped to both Pep1 and Pep2, representing a case that multiple PSMs can be included in the rescuing process. Highlighted Scan1 and Scan4 represent valid identifications that pass a threshold of confidence. (B) All peptides are scored and prioritized by the "Bayesian average" rating method. In this example, Pep2 receives the best Bayesian average score, and Pep3 is the lowest rated peptide. (C) IDBoost sifts through prioritized peptides and rescues unidentified spectra.

Including multiple PSMs per spectrum in the rescuing process enables the rescue of secondary peptide identifications, but this dramatically increases the number of PSMs to be processed. Many of these PSMs are assigned low database search scores and are not likely to be confident identifications. To reduce processing time, IDBoost provides a

configurable filter to exclude PSMs of low score. For example, throughout this study, PSMs with MyriMatch MVH scores lower than 10 were excluded.

It should be noted that IDBoost only increases spectral identifications mapping to currently identified peptides, while peptide and protein identifications remain unchanged unless a spectral count-based filter is applied in protein inference. IDBoost does not remove originally identified PSMs. If a PSM is rescued for an identified spectrum, both the original PSM and the rescued PSM will be presented, implying that the rescued peptide is better supported by a cluster of spectra than the original identification. The IDPicker tool in which IDBoost is implemented provides a graphical user interface (GUI) to present both rescued and originally identified PSMs. It also displays database search scores and offers a spectrum viewer to visualize peptide-spectrum matches, enabling manual validation of ambiguous identifications if multiple peptides are assigned to a single spectrum.

II.2.4 Bayesian Average Score

As illustrated in Figure 9B, all peptides in a cluster are prioritized for rescue by “Bayesian average” scores. “Bayesian average” is a rating method to calculate the mean of a set of data that is consistent with Bayes’ theorem. Given a set of options rated by a number of voters, instead of simply calculating the average rating of an option, the “Bayesian average” method incorporates the number of votes into the calculation, generating a weighted average score. As a result, options with more votes receive Bayesian average scores closer to their unweighted arithmetic average. In contrast, when there are few votes, the rating of an option will be a weighted average (a Bayesian

average score) that is closer to the average rating of all options. In this study, each peptide is an option, and spectra in a cluster are voters. The database search scores of PSMs represent ratings between spectra and peptides. The “Bayesian average” score of a peptide is computed to be:

$$\bar{x} = \frac{C * m + \sum_{i=1}^{n} x_i}{n + C}$$

where x is a database search score assigned to this peptide, n is the number of spectra mapped to this peptide, m is the mean of database search scores taken over all PSMs in this cluster. C is a weighting constant that should be a large number and is proportional to the size of the dataset. Here I use the maximal number of spectra assigned to a peptide in this cluster, i.e., the maximal votes of an option in the dataset, multiplied by 10 to keep it a large number. The Bayesian average reflects how peptides are scored in database search in relation to each other. A peptide identified by a relatively large number of spectra receives a Bayesian average score close to its unweighted average. In converse, the Bayesian average score for peptides with a relatively small number of spectra tends to gravitate towards the average rating of all PSMs.

To prioritize peptides in a cluster, the Bayesian average scores are normalized to percentiles. For example, a Bayesian average score of 0.99 means this peptide is scored better than 99% of other peptides in a cluster. If the same set of spectra is searched by multiple database search engines, the Bayesian average scores are computed separately for each analysis, and then summed together to rank peptides. In this case, peptides shared by multiple search engines are more believable and will receive higher Bayesian average scores. IDBoost exports a tab-delimited text file to report rescued PSMs and their Bayesian average scores.

During the method development, I also considered several other voting methods such as “Borda count” and “Condorcet method”. I decided to choose the Bayesian average rating because it allows weighting voters and its success has been proved in many user reviewing systems.

II.3 Data Sources

Several datasets were used to demonstrate the utility of IDBoost (see Table 2). Binary spectral data present in the raw files were converted to mzML (Deutsch 2008) format using MSConvert tool of the ProteoWizard (Kessner et al. 2008) library. The MyriMatch tool searched each file against a protein database that contained sequences in both forward and reverse orientations for estimation of protein identification error rates. Search results were processed by IDPicker for peptide validation and protein assembly. Throughout this study, IDPicker was configured to derive PSM score thresholds to yield a 5% FDR. Detailed configurations of MyriMatch and IDPicker are given in Appendix A.

“Synthetic Orbi” Dataset

This dataset was previously used to test a phosphorylation site localization algorithm and the experimental description was published (Savitski et al. 2011). In brief, 180 peptides with positional phosphosite isomers were synthesized and pooled to five mixtures, such that no phosphorylation site isomers were present in any one mixture. Mixtures were analyzed on a Thermo Fisher LTQ-Orbitrap hybrid mass spectrometer in which peptides were fragmented by CID.

Dataset	# of files (sample x rep)	Average # of MS/MS scans	Databases used for search
Rescue of Phosphopeptide Spectra			
Synthetic Orbi	5 x 1	1598	IPI.HUMAN.v3.79
pTyr LTQ	1 x 3	18550	IPI.HUMAN.v3.79
Rescue of Spectra in Comparative Analysis			
Yeast LTQ	1 x 3	26151	SGD.orf_trans_all+UPS1
Yeast_UPS1 LTQ	5 x 3	26148	SGD.orf_trans_all+UPS1
Rescue of Spectra in a Variety of Data			
UPS1 LTQ	3 x 3	24937	SGD.orf_trans_all+UPS1
UPS1 Orbi	3 x 3	10935	SGD.orf_trans_all+UPS1
Yeast LTQ	3 x 3	24948	SGD.orf_trans_all+UPS1
Yeast Orbi	3 x 3	12464	SGD.orf_trans_all+UPS1
Yeast MudPIT LCQ	19 x 6	2961	SGD.orf_trans_all

Table 2. Experimental datasets for the evaluation of IDBoost.

"pTyr LTQ" Dataset

A human epithelial carcinoma cell line (A431) (ATCC, Manassas, VA) was cultured in 150 mm culture dishes in improved MEM (Invitrogen-GIBCO, Auckland, NZ) supplemented with 10% fetal bovine serum (Atlas Biologicals, Fort Collins, CO) at 37°C in 5% CO_2. A431 cells were grown to ~60-70% confluency prior to treatment. Cells were serum-starved (18 hrs), followed by treatment with 30 nmol epidermal growth factor (EGF)(Cell Signaling Technology, Danvers, MA) for 30 min. Cells were harvested on ice with Mg and Cl-free PBS supplemented with a phosphatase inhibitor cocktail (1 mM sodium fluoride, 10 mM β-glycerophosphate, 1 mM sodium molybdate, and 1 mM activated sodium orthovanadate – individual components purchased from Sigma (St. Louis, MO)), pelleted by centrifugation at ~250 x g, flash-frozen and stored at -80°C.

The phosphotyrosine enriched dataset was generated by enriching phosphotyrosine peptides from tryptic digests of cell lysates as previously decribed (Rush et al. 2005) except that cells were lysed in 50:50 (v/v) acetonitrile and 50mM ammonium bicarbonate prior to in-solution trypsin (Promega, Madison, WI) digestion and samples were pY enriched using 4G10 antibody (Millipore, Billerica, MA). LC-MS/MS and MS3 analyses were performed on a Thermo Fisher LTQ Velos (San Jose, CA) mass spectrometer equipped with an Eksigent Nano-1D Plus HPLC and AS-1 autosampler (Dublin, CA). Peptides were separated on a 100 μm × 11 cm fused silica capillary column (Polymicro Technologies, LLC., Phoenix, AZ) and 100 μm x 6 cm fused silica capillary precolumn packed with 5 μm, 300 Å Jupiter C18 (Phenomenex, Torrance, CA). Liquid chromatography was performed using a 95 min gradient at a flow rate of either 400 or 600 nL min−1 using a gradient mixture of 0.1% (v/v) formic acid in water (solvent A) and 0.1% (v/v) formic acid in acetonitrile (solvent B). Briefly, a 15 min wash period (100% solvent A) was performed followed by a gradient to 98% A at 15 min (1.2 μl min-1) and eluent was diverted to waste prior to the analytical column using a vented column set up similar to that previously described (Licklider et al. 2002). Following removal of residual salts, the flow was redirected to flow through the analytical column and solvent B increased to 75% over 35 minutes and up to 90% in 65 minutes. The column was re-equlibrated to 98% solvent A for 10 minutes after each run. MS/MS peptide spectra were acquired using data-dependent scanning in which one full MS spectrum was followed by 5 MS/MS spectra. A data-dependent scan for the neutral loss of phosphoric acid or phosphate resulted in acquisition of an MS/MS/MS of the neutral loss ion.

"Yeast LTQ", "Yeast Orbi", "UPS1 LTQ", "UPS1 Orbi" and "Yeast_UPS1 LTQ" Datasets

These datasets are publically available for download from Proteome Commons website (https://www.proteomecommons.org) and the experimental details are available in the original publication (Paulovich et al. 2010). Yeast lysate was reduced by dithiothreitol (DTT), alkylated by iodoacetamide and digested by trypsin. Both yeast and UPS1 (Sigma UPS1, Sigma-Aldrich, St. Louis, MO) were analyzed on LTQ and LTQ-Orbitrap instruments. The "Yeast_UPS1" data represents a mixture of yeast and spiked UPS1 in five different concentrations: 0.24, 0.67, 2.54, 6.7 and 20 fmol/μl. This sample was analyzed on a LTQ instrument.

"Yeast MudPIT LCQ" Dataset

This dataset was published by Arnett et al. and the experimental details were described in the original publication (Arnett et al. 2008). In brief, Weil lab at Vanderbilt University collected spectra from 19 MudPIT experiments to study yeast Mot1p protein-protein interactions, in which immunopurifications of Mot1p-interacting proteins were performed using multiple antibodies. Each pull-down was subjected to MudPIT analysis with six fractions and analyzed on a Thermo LCQ Deca XP Plus mass spectrometer.

II.4 Results and Discussion

To establish the effectiveness of IDBoost, I first evaluated the method using two phosphorylation datasets. I show that by encompassing search results from similar spectra, IDBoost achieved high accuracy in rescuing correct identifications. Next, I demonstrate

the use of IDBoost to enhance the recognition of differentially expressed proteins in comparative analysis. I then demonstrate IDBoost performance in a variety of datasets. These tests established IDBoost as an effective and robust method to rescue confident spectral identifications.

II.4.1 Rescue of Phosphopeptide Spectra to Resolve Phosphosite Localization Ambiguity

Once similar spectra are clustered together, the rescuing process starts from the peptide assigned the best Bayesian average score. To ensure that the correct peptides are rescued, "Bayesian average" method should be able to score true peptides more highly than random matches. This is a less serious problem in database searches to produce inventories, because a set of spectra in a cluster often maps to a single peptide. However, this becomes more complicated in phosphopeptide searches, in which spectra may be identified to the same peptide sequence with different phosphorylation sites, i.e., phosphosite isomers. In database search, phosphosite isomers often score very similarly even though true peptides generally receive better scores than false isomers. I expect the "Bayesian average" method in IDBoost to rate a true peptide sequence more highly than one with misplaced modification site. Incorporating search results from similar spectra could thus reduce phosphosite localization ambiguity.

To evaluate the effectiveness of the "Bayesian average" method, I used a synthetic phosphopeptide dataset in which peptide sequences and phosphorylation sites are known (Savitski et al. 2011). All spectra were searched using MyriMatch against an IPI human protein database and post-processed by IDPicker.

I first examined all spectra that were assigned known sequences and correct PTM locations within their top 5 PSMs. Among all five raw files, 1678 of 7802 spectra were assigned to the correct synthetic sequences, while only 945 of them were confident identifications that passed the 5% FDR threshold. After running IDBoost, this number increased to 1348, augmenting sensitivity from 56% to 80%. Next, I evaluated the accuracy of rescued PSMs. A total of 1148 spectra (945 matched to synthetic sequences) were confidently identified in the original analysis, and 586 additional PSMs were rescued. As shown in Figure 10A, 69% of rescued spectra were correctly assigned to synthetic peptides with known PTM locations. Only 2% were false rescues that were originally assigned to the correct sequences, but then rescued to different ones. A close look at these false rescues revealed that they all were phosphorylation site isomers. The remaining 29% of rescued PSMs were associated with peptides that were not included in synthetic mixtures and thus are labeled “Unknown.”

A sub-pie-chart in Figure 10A shows the proportion of top-ranked and secondary PSMs that were correctly rescued. 54% of rescued PSMs were correct top-ranked hits that failed to pass the FDR threshold. 10% correctly rescued PSMs were originally assigned wrong peptide sequences and 5% were assigned to phosphosite isomers. The result shows that both secondary phosphosite isomers and spurious random matches can be corrected through the rescuing process.

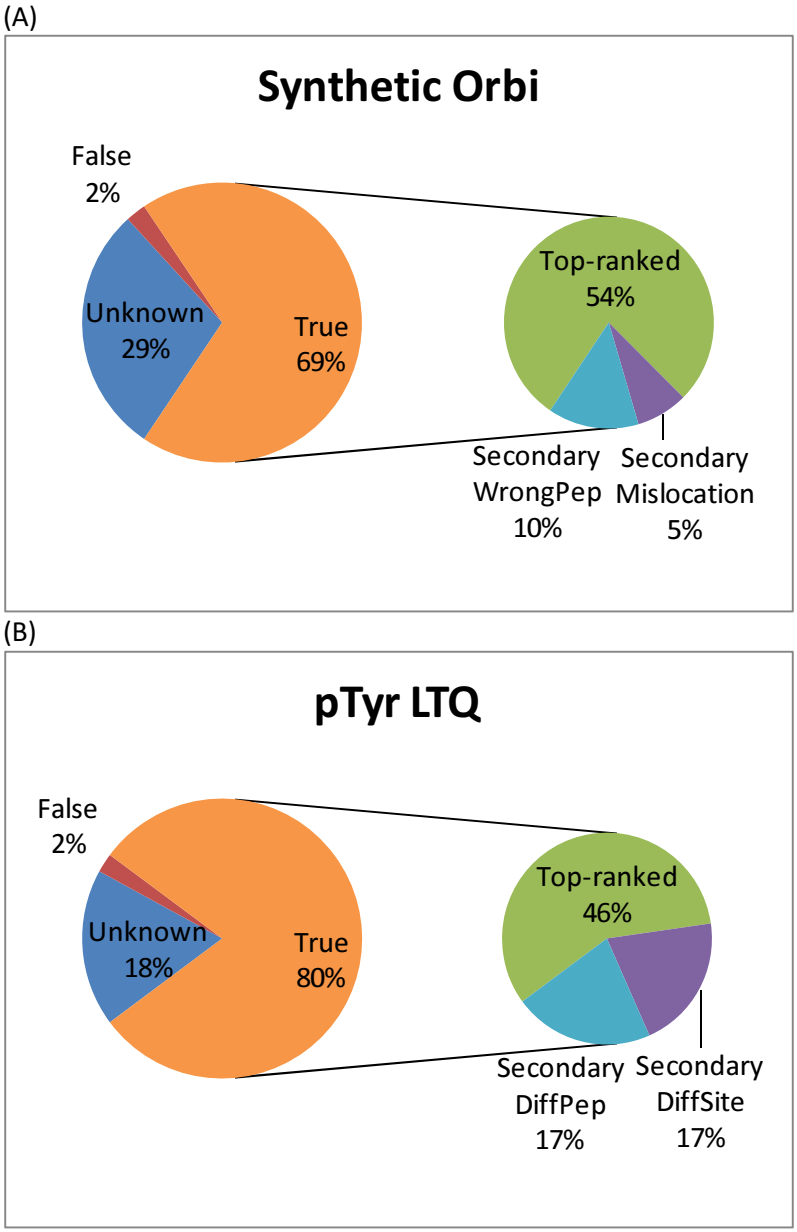

Figure 10. Analysis of rescued PSMs in phosphorylation studies. In each panel, the left pie chart shows the accuracy of rescued PSMs and the right sub-pie-chart represents a more detailed examination of correctly rescued PSMs. (A) Rescued PSMs from a synthetic phosphopeptide dataset. (2) Rescued PSMs from a phosphotyrosine enriched dataset. The result shows that IDBoost is able to recognize correct phosphosite isomers for rescue.

This test established the "Bayesian average" method as an effective way to prioritize peptides in a cluster. Among 1678 spectra that were assigned correct sequences

within their top 5 PSMs, 70% contained phosphosite isomers, implying that a large number of spectra were mapped to multiple closely scored isomers during the rescuing process. The result indicates that the "Bayesian average" method is able to score correct sequences better than their phosphosite isomers, enabling rescue of these correct sequences rather than their isomers. Moreover, the result can be applied to resolve the ambiguous phosphosite localization. If a phosphosite isomer is rescued, it implies that this phosphorylation site is better supported than the original assignment by a cluster of similar spectra. In addition, this test illustrated that IDBoost is effective, even in datasets that contain a single analysis of each sample and where the MS/MS analyses employed dynamic exclusion to reduce repeated sampling of each peptide.

It should be noted that in complex biological samples a single peptide may be singly phosphorylated at multiple locations, i.e., multiple positions all may be correct identifications. If similar spectra are produced by these isomers, the one with stronger identification evidence (better database search scores or more votes) may be scored superior to the correct position, thus rescuing a false positioning. Most likely this happens with phosphopeptides that produce similar fragment ions. In this case, IDBoost provides alternative interpretations for further manual validation.

Next, I tested IDBoost performance using a real-world biological sample. The Jim Ayers Institute at Vanderbilt University collected three technical replicate runs of a human epithelial carcinoma cell lysate after enriching phosphotyrosine peptides with 4G10 antibodies. The samples were analyzed on a Thermo Fisher LTQ Velos mass spectrometer. After MyriMatch and IDPicker analysis, 3327 spectra were confidently identified, counting all spectra without regard to phosphorylation status. In this test, 1050

PSMs were added via IDBoost. The number of spectra assigned to phosphotyrosine peptides before and after running IDBoost was 1967 and 2512, respectively (a 28% increase).

To estimate the accuracy of rescued PSMs, I considered a PSM as being rescued properly if its peptide contained a phosphotyrosine modification. In converse, PSMs that contained phosphotyrosine in original assignments, but which were rescued to non-phosphotyrosine peptides were treated as false rescues. As shown in Figure 10B, 80% of rescued PSMs were proper rescues and 2% were false. The other 18% of PSMs that did not fall into these two categories were labeled "Unknown." A close examination of the proper rescues showed 46% of rescued PSMs were top-ranked. 17% of spectra were rescued to secondary PSMs that were phosphosite isomers of top-ranked peptides. 17% of spectra were assigned different peptide sequences in the original analysis and were rescued to phosphotyrosine peptides. The result indicates that a large number of spectra assigned to phosphotyrosine peptides can be rescued. In addition, ambiguous phosphosites can be further evaluated in the context of a cluster of similar spectra. Bayesian average scores assigned to phosphosite isomers imply which phosphosite is better supported by these spectra than the other.

II.4.2 Rescue of Spectra in Comparative Analysis

In spectral count-based comparative analysis, differentially expressed proteins are determined by comparing the number of spectra observed for these proteins between pairs of cohorts. Generally, a larger average count difference yields a more significant result in statistical testing. To test if IDBoost helps to enhance spectral count

differentiation, I used a standardized dataset from the National Cancer Institute Clinical Proteomic Technologies Assessment for Cancer (CPTAC) Study 6 (Paulovich et al. 2010), in which a mixture of 48 human proteins (Sigma UPS1, Sigma-Aldrich, St. Louis, MO) was spiked into the yeast reference proteome at different concentrations: A: 0.24, B: 0.67, C: 2.54, D: 6.7, E: 20 fmol/μl and no spikes. After MyriMatch database search, five IDPicker analyses were performed to compare proteins between each concentration group and yeast sample that has no spikes (i.e., group A vs. yeast, group B vs. yeast etc.). I calculated the spectral count difference for each protein between yeast sample and the sample spiked with human proteins. The average differences for 48 spiked proteins and background proteins were compared before and after running IDBoost. In the presence of high concentrations of spiked proteins, the average differences for those proteins were much larger than for background proteins (see Figure 11, groups D and E), while the differences became less distinguishable for samples with low concentration spikes (see Figure 11, groups A and B). For all groups, IDBoost enlarged the spectral count differences for spiked proteins, with marginal effects on background proteins. Since these differences are the fundamental evidence for most statistical tests to determine differentially expressed proteins, running IDBoost improves sensitivity in differential proteomics by allowing the spiked proteins to be better recognized.

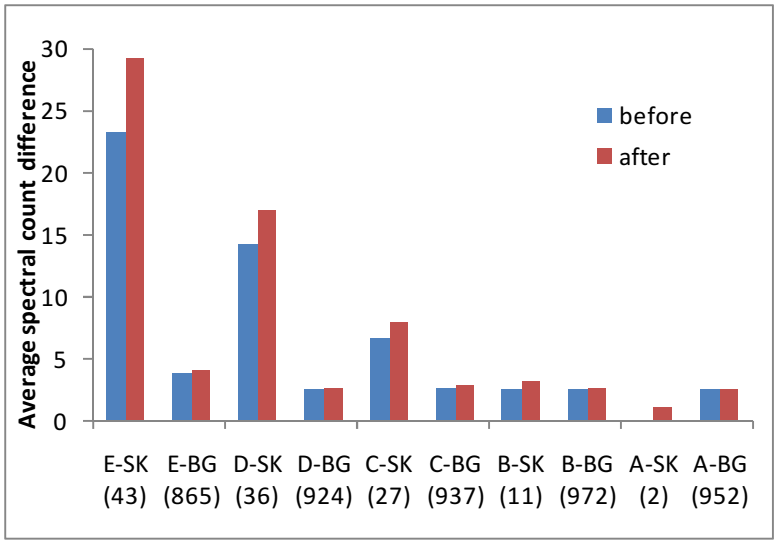

Figure 11. Impact of IDBoost on recognition of differentially expressed proteins in comparative analysis. 48 human proteins were spiked into a yeast proteome with decreased concentrations from group E to A. The spectral count differences between two samples, yeast and yeast with spikes, were calculated for each protein. Here I examine the average number of differences for two groups of proteins, 48 spiked proteins (SK) and the background yeast proteins (BG). The numbers of identified proteins are enclosed in parentheses. In all tests, IDBoost enhances the spectral count differences of differentially expressed proteins (spiked proteins) in comparative analysis.

This method is particularly valuable for samples with low concentration of differentially expressed proteins. In Figure 11, for example, fewer spiked proteins were identified as their concentrations decreased. The high concentration samples E and D may benefit less from IDBoost due to the fact that the spectral count differences for spiked proteins were already much larger than background proteins. However, the comparative analysis may be improved by the use of IDBoost for sample C and B in which the differences of spiked proteins were close to those of background proteins. In these cases, by increasing the spectral count differences for spiked proteins, IDBoost allows the spiked proteins being selected more confidently in statistical analysis. Sample A was intentionally spiked with too low a concentration for most differences to be

observed. Only 2 of 48 spiked proteins were identified, and no differences were found between experiments.

II.4.3 Rescue of Spectra in a Variety of Datasets

I first tested IDBoost performance using two sample mixtures collected for CPTAC Study 6 (Paulovich et al. 2010). The Sigma UPS1 sample (Sigma-Aldrich, St. Louis, MO) is a defined mixture that contains 48 human proteins in equimolar concentrations. The yeast sample is a protein extract of *Saccharomyces cerevisiae*, representing a highly complex biological proteome. Both samples were prepared by the NIST and shipped to the CPTAC sites. I selected nine files for each sample (triplicates from three instruments) collected from two instrument platforms: a high resolution Thermo Fisher LTQ-Orbitrap and a lower resolution LTQ linear ion trap mass spectrometer. All spectra were searched using MyriMatch and post-processed by IDPicker. For each sample, I ran IDBoost against either all nine files collected from an instrument type or within the three files collected from a particular instrument.

Figure 12 shows the number of spectral identifications before and after running IDBoost, along with the percent of gained spectra in each analysis. This figure demonstrated IDBoost performance in a variety of sample complexities and instrumentation. First, simple mixtures (UPS1) benefit more from the rescue process than do complex samples (yeast). The percent of gained spectra varied from 17% to 52% for the UPS1 sample, and these gains were always higher than for yeast (below 15%). Second, data from low resolution instruments tend to gain more identifications than those from high resolution instruments. In both UPS1 and yeast samples, the proportions of

rescued spectra were higher in LTQ data than those in Orbitrap runs, probably because the high resolution data yielding more confident IDs in MyriMatch search, leaving fewer spectra for rescue. Third, IDBoost shows enhanced performance for datasets with more replicates, even when they come from different instruments. Processing all files together yielded more rescued spectra than processing each instrument set separately.

In terms of running time, IDBoost spent around 1 minute to process each set of triplicates from an individual instrument for UPS1 data and 2 minutes for yeast data on a Dell Optiplex 745 computer with an Intel Core 2 Duo 6400 processor and 3 GB of RAM. When processing all nine files together, IDBoost spent 6 minutes on UPS1 LTQ data and 2 minutes on UPS1 Orbi data. It took about 8 minutes to process all nine files for yeast LTQ and Orbi data.

I also tested IDBoost on a large-scale study using MudPIT technology (Arnett et al. 2008). 63195 of 337602 spectra were identified by MyriMatch and IDPicker analysis. After running IDBoost, the spectral identifications increased 26% to 79709. IDBoost spent 24 minutes to process all 116 files conjointly.

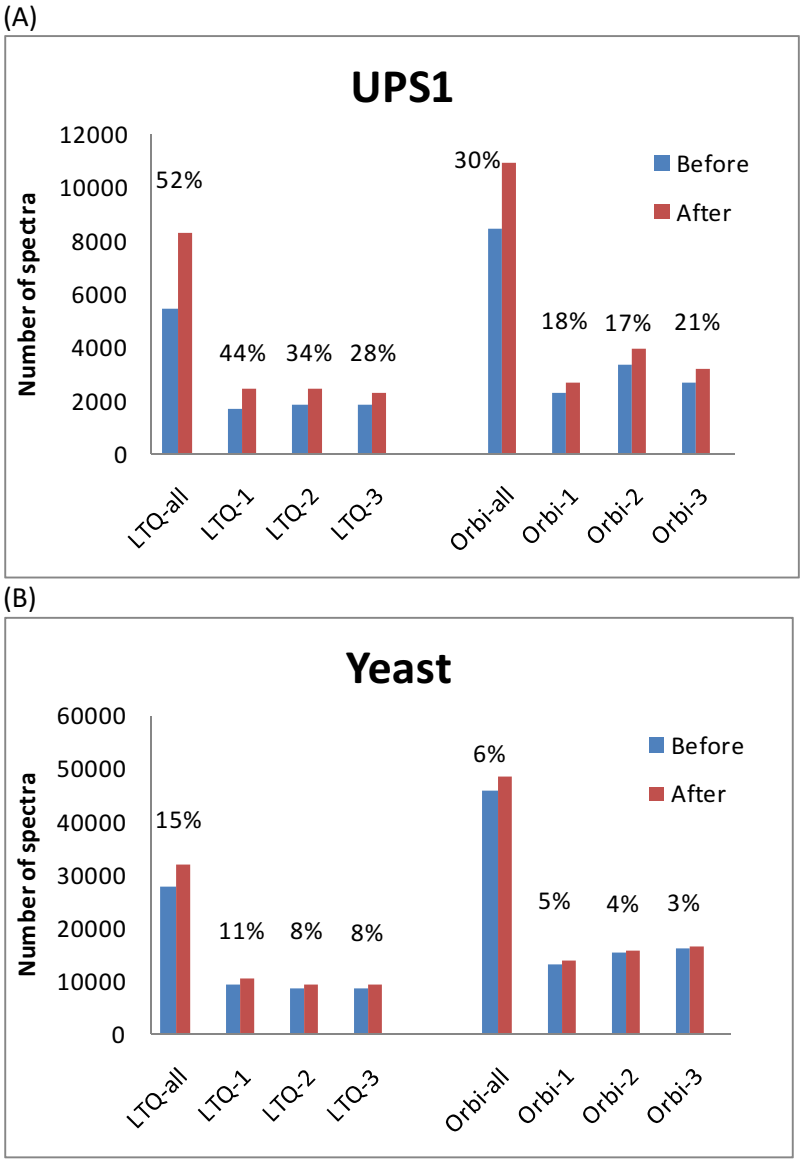

Figure 12. IDBoost performance in a variety of datasets. UPS1 (A) and yeast (B) samples, each with three technical replicates, were analyzed on three individual instruments in two instrument platforms. Search results from either each instrument (3 files) or all 9 files from an instrument platform were processed using IDBoost. The number of spectra before and after running IDBoost is presented. The proportion of gained spectra is reported for each analysis.

II.5 Conclusion

I presented a method to rescue spectral identifications and correct database search errors through spectral clustering. I demonstrated the use of IDBoost in phosphorylation

studies and comparative analysis. I tested the effectiveness of IDBoost in a variety of datasets. Experiments with many replicates and low accuracy data tend to benefit most from the use of IDBoost. IDBoost provides an easy and fast way to expand confident spectral identifications based on existing analysis with no requirement of additional identification steps. Its use is not limited to particular search engines or post-processing tools and thus it can be integrated into established proteomics data analysis workflows. The IDPicker GUI in which IDBoost is embedded enables visualization and manual validation of rescued identifications.

To cluster similar spectra, IDBoost computes a dot product for each pair of spectra. This method has been proved to be effective in most cases but is highly affected by major peaks. In the future, more robust methods such as the scoring system in Pepitome can be implemented to replace the dot product for spectra similarity comparison.

CHAPTER III

SCANRANKER: QUALITY ASSESSMENT OF TANDEM MASS SPECTRA VIA SEQUENCE TAGGING

III.1 Introduction

A large number of high quality spectra remain unidentified after database search due to modifications, incompleteness of protein databases, constrained search parameters and the deficiencies of the scoring methods in database search tools. These spectra often represent meaningful biological information and are potentially identifiable with alternative approaches such as blind modification search and *de novo* sequencing (Ning et al. 2010). An automated spectral quality assessment tool helps to ameliorate these problems. It can be used to find unidentified high quality spectra for subsequent analysis and helps to select high quality spectra for *de novo* sequencing.

Mass spectrometry has become a method of choice to characterize cross-linked proteins (Leitner et al. 2010). The identification of cross-linked peptides, however, is quite a daunting job due to the overwhelming number of possible matches and the difficulty of interpreting spectra from cross-linked peptides. Although several bioinformatics tools have been developed to relieve theses difficulties, manual confirmation of cross-linked peptides is generally necessary. A spectral quality assessment tool could facilitate this process by providing a ranked list of spectra for manual interpretation.

The spectral quality score can also be used in the process of peptide assignment validation. In database search, software tools usually assign different scores to measure the match between spectrum and peptide (e.g., XCorr from Sequest and IonScore from Mascot), which are subsequently used in statistical analysis to estimate FDR. The spectral quality score could become an additional score in this process, because high quality spectra are more likely to produce confident peptide identifications.

The scoring methods in sequence tagging algorithms are applicable for quality assessment of tandem mass spectra. A high quality spectrum of a peptide is expected to contain a series of consecutive fragment ions corresponding to peptide bond breakages (Tabb et al. 2006). These fragments provide a basis for partial sequence inference that result in multiple tags with good scores. Conversely, if no sequence tags can be inferred from a spectrum, it is unlikely that the spectrum will produce a high score in database search. Sequence tagging is a robust approach for spectral quality assessment because even modified or mutated peptides can produce consecutive fragment ions. Recently, we developed a novel sequence tagging algorithm, DirecTag (Tabb et al. 2008), which demonstrated superior accuracy in comparison to existing sequence tagging tools. In this work, I explore the use of DirecTag along with other metrics for spectral quality assessment.

Several spectral quality assessment tools have been developed in recent years. Pioneering work by Bern et al. (2004) predicted spectral quality based on a set of handcrafted features Other studies by Xu et al. (2005) as well as by Salmi et al. (2006) reported a quadratic discriminant function and a random forest classifier to separate good and bad spectra, respectively. Na & Paek (2006) proposed a cumulative intensity

normalization method for quality assessment, while Flikka et al. (2006) tested several machine learning classifiers in data from three different mass spectrometers, recognizing that the performance of classifiers is greatly affected by the type of instrument. More recently, Nesvizhskii et al. (2006) developed QualScore, which produces accurate results to find unassigned good spectra after database search. In these prior studies, the proposed methods were usually evaluated based on their performance in removing low quality spectra and recovering unassigned high quality spectra. In fact, quality assessment tools are useful for a wide variety of applications that have not previously been demonstrated. These tools may help to prioritize spectra for *de novo* sequencing and cross-linking analysis, which are usually very time-consuming processes relying heavily on manual inspection. Besides, since high quality spectra are more likely to produce confident identifications in database search, the quality assessment tools can also be used for quality control of datasets in large-scale proteomic studies.

In this work, I present ScanRanker, a new software tool that evaluates spectral quality via sequence tagging. I evaluate ScanRanker using a variety of datasets from multiple instrument platforms with different sample complexities. I demonstrate that ScanRanker can be used both to recognize high quality spectra that fail identification and to remove low quality spectra prior to database search. In addition, I demonstrate several applications of spectral quality score that are not explored in existing publications. I show that ScanRanker scores can be used to predict the richness of identifiable spectra among LC-MS/MS runs in an experiment. I demonstrate the use of ScanRanker scores in the process of peptide assignment validation. I also demonstrate that ScanRanker helps to select high quality spectra for *de novo* sequencing and cross-linking analysis.

III.2 Algorithm

III.2.1 Overview

ScanRanker makes use of the DirecTag algorithm to infer sequence tags from tandem mass spectra. It then computes a quality score for each spectrum on the basis of three tag-based scoring metrics: “BestTagScore”, “BestTagTIC” and “TagMzRange”. ScanRanker accepts spectra in mzML, mzXML and MGF file formats via use of the ProteoWizard library. Several proprietary formats, such as Thermo RAW files and Bruker YEP files, can also be directly processed with no required installation of vendor-supplied software libraries (a detailed list of supported formats is available at http://proteowizard.sourceforge.net/docs.html). ScanRanker can be executed in both Microsoft Windows and Linux systems, though native support for vendor formats requires use of Windows. A GUI was created in C#/.NET for Windows users. A helper program, IonMatcher, was also developed to visualize ScanRanker results and enable interactive manual inspection of peptide-spectrum matches. The source code and executable versions of ScanRanker are available from http://fenchurch.mc.vanderbilt.edu.

The screenshot of the ScanRanker GUI is shown in Figure 13. The ScanRanker GUI contains three major parts: "Spectral Quality Assessment", "Spectral Removal" and "Spectral Recovery". The "Spectral Quality Assessment" feature controls parameters for running sequence tagging by DirecTag. It writes out a metrics file, which can be used later for "Spectral Removal" and “Spectral Recovery”. If the charge state of a spectrum is not determined (for example, LTQ data), a spectral quality score will be assessed for each charge state, and the highest quality score will be retained.

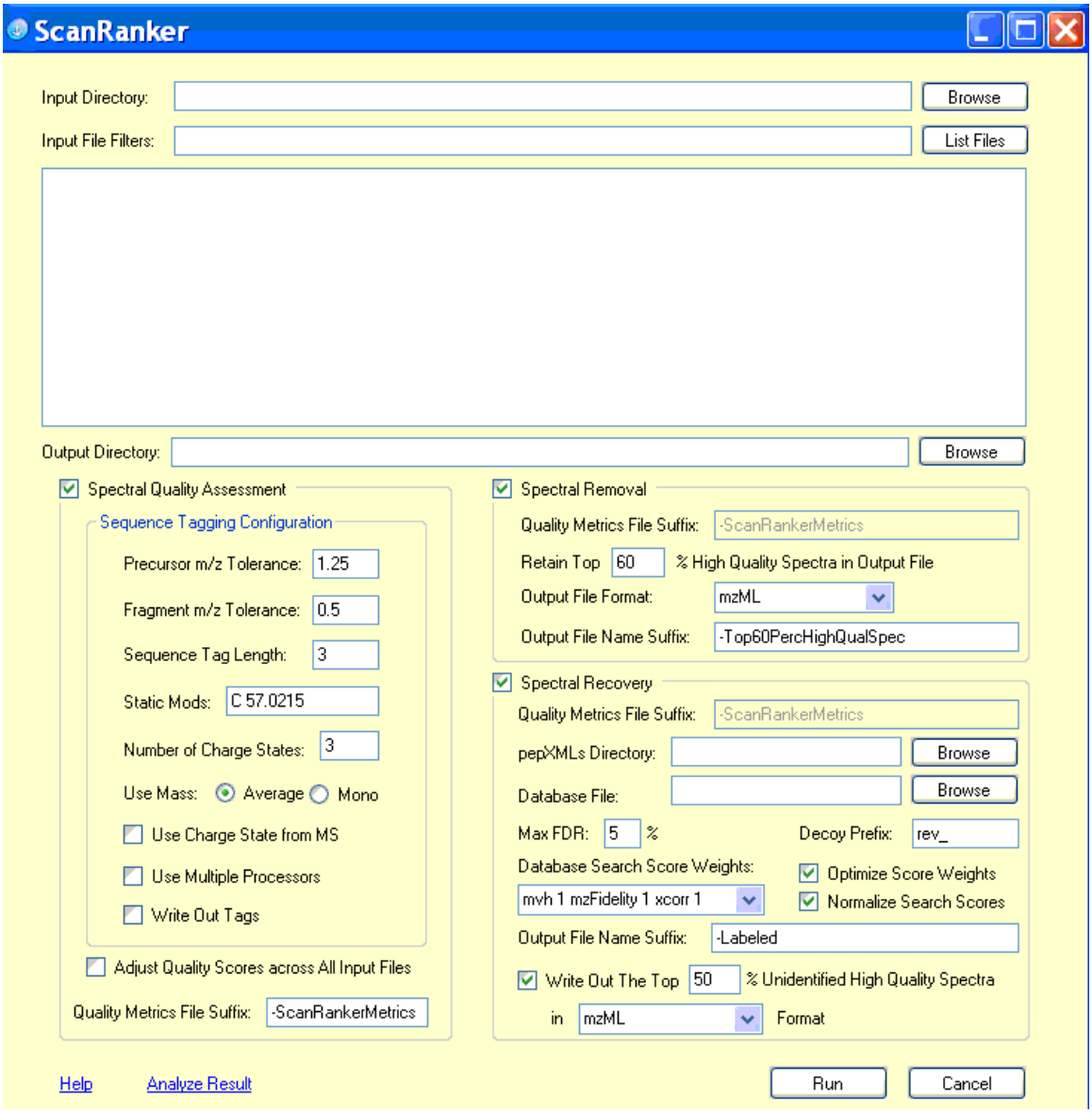

Figure 13. A screenshot of ScanRanker GUI.

The “Spectral Removal” feature generates a subset of high quality spectra in mzML, mzXML, MGF or MS2 format, which can be used for more intensive searches. The "Spectral Recovery" feature makes use of "idpQonvert" module in IDPicker software to determine which spectra are identified. Based on the idpQonvert result, it adds a label (1 or 0) to each spectrum in a metrics file to indicate whether the spectrum is identified

by IDPicker. The corresponding peptides and proteins of identified spectra will also be included in the metrics file. ScanRanker generates unidentified high quality spectra for further analysis such as *de novo* sequencing and cross-linking analysis.

The screenshot of the IonMatcher GUI is shown in Figure 14. IonMatcher reads a spectrum file and a metrics file to allow manual inspection of spectral quality. More importantly, it enables interactive validation of peptide-spectrum matches. If a metrics file is generated by "Spectral Recovery", the identified peptide sequence will be displayed in a data table. Clicking a row in the table brings up four panels: annotation panel, fragmentation panel, spectrum panel and *de novo* sequencing panel. The peptide sequence in annotation panel can be modified interactively to exam the match between a modified sequence and the spectrum. Cross-correlation scores are reported for each sequence. The fragmentation panel displays *m/z* values of selected fragment ion series in which matched ions are bold highlighted. The spectrum panel shows matched ions and fragmentation ladders.

If no peptide was assigned to a spectrum in database search, potential interpretations of the spectrum can be inferred using PepNovo, a state-of-the-art *de novo* sequencing tool developed at University of California, San Diego (UCSD). Inferred peptide sequences can be copied to annotation panel for manual validation. It should be noted that PepNovo program is not included in the ScanRanker package. To enable the *de novo* sequencing function, please download PepNovo at http://proteomics.ucsd.edu/index.html and copy all files to the \ScanRanker-installation-directory\PepNovo folder. Copyright and License information of PepNovo are available in UCSD website.

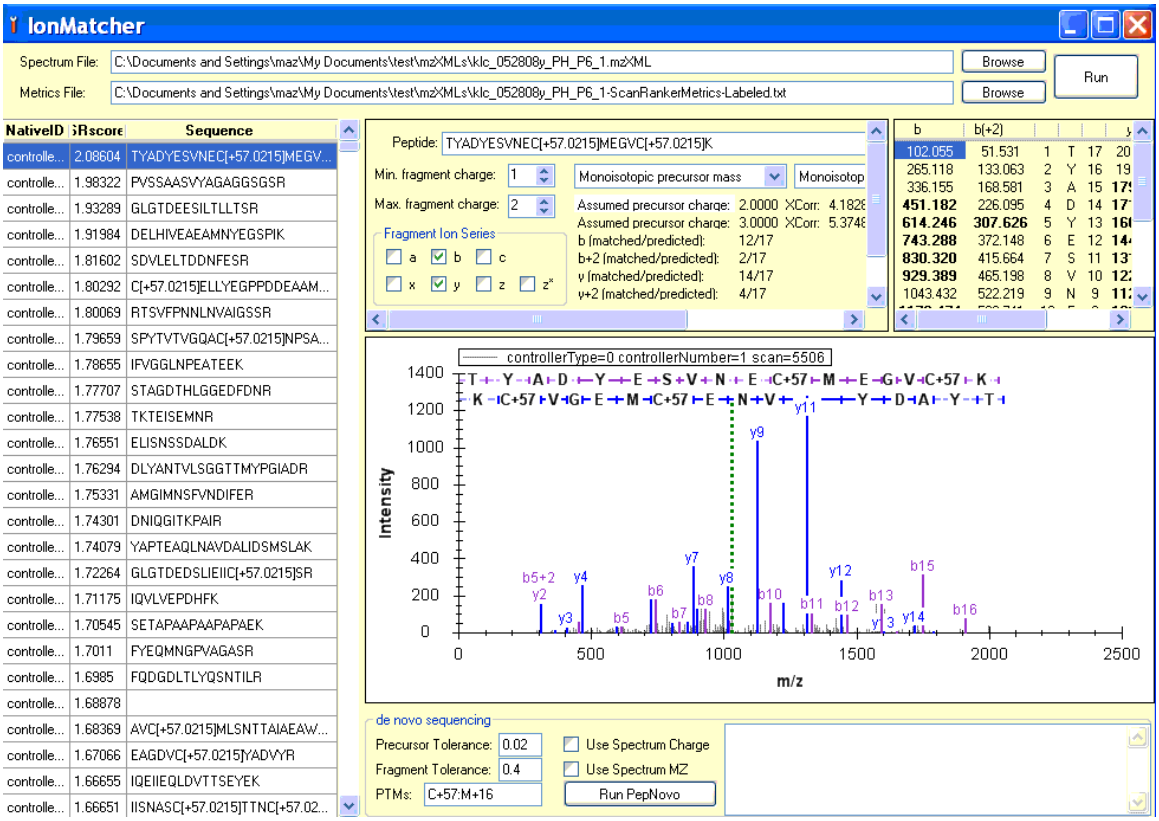

Figure 14. A screenshot of IonMatcher GUI.

III.2.2 BestTagScore Subscore

DirecTag evaluates sequence tags on the basis of peak intensity, *m/z* fidelity and complementarity. Each tag is assigned a p-value to represent the probability that a better score would have resulted by chance. Here I made use of the score of the top ranked tag as the "BestTagScore" subscore for spectral quality assessment. Spectra that are capable of generating high quality tags are more likely to be good spectra.

III.2.3 BestTagTIC Subscore

To infer sequence tags, DirecTag constructs a graph comprising nodes representing peaks and edges representing pairs of peaks that are separated by amino acid

masses. DirecTag seeks out consecutive edges in this graph to enumerate sequence tags. For example, a set of four connected nodes in the graph may constitute a tag of three amino acids. Each node in a spectrum graph is associated with a peak intensity value. The "BestTagTIC" subscore sums up peak intensities of the top ranked tag. A high quality spectrum is expected to have a higher "BestTagTIC" subscore than low quality ones in a dataset. Spectra that are higher in intensity are more likely to produce tags of high TIC.

III.2.4 TagMzRange Subscore

Each inferred tag corresponds directly to a series of fragments in a tandem mass spectrum. The "*m/z* range" of a tag is the *m/z* distance that extends from the first peak to the last peak of the tag. By examining all enumerated tags, the "TagMzRange" subscore describes the widest range of *m/z* values for a spectrum that is spanned by tags. For a spectrum generating many tags, the "TagMzRange" subscore is equal to the *m/z* range between the lowest *m/z* peak and the highest *m/z* peak across all enumerated tags minus any *m/z* areas that are not spanned by tags. If tags can be generated from a wide *m/z* range in a spectrum, it is more likely that this spectrum will be identifiable by computational tools.

III.2.5 Spectral Quality Score

Three subscores are subjected to logarithmic transformation and normalized before generating a final quality score. The normalization of each subscore is performed by subtracting the mean of subscores in that dataset, and then divided by the interquartile range of these subscores. Spectra with no inferred tags or the best scored tags exceeded

the threshold specified in configuration file, usually 10-20% of spectra in a dataset, are considered as low quality spectra and are excluded in the calculation of mean and interquartile range. ScanRanker computes the average of three normalized subscores as the final quality score. Multiple LC-MS/MS runs, such as MudPIT or gel band runs, can be optionally grouped together as a single experiment, for which the mean and interquartile range of subscores across all datasets will be used for normalization.

During developing the scoring method, I also attempted to use logistic regression and support vector machine based models to generate quality scores. These models can handle a large number of variables, so other attributes such as the number of peaks in a spectrum, total ion intensity and the ratio of strong and weak peaks in a spectrum can be incorporated into the scoring system. However, I found the proposed method with three variables can achieve almost the same performance as using more variables in sophisticated models. Therefore, only three most discriminating features were retained for quality assessment here.

III.3 Data Sources

The evaluation of the ScanRanker algorithm employed several datasets collected from different instrument platforms (see Table 3). The configurations of ScanRanker and other software tools are given in Appendix A. Instrument raw files were converted to mzXML format using the MSConvert tool of the ProteoWizard library. DTA format files required for Sequest search were extracted from the mzXML files using mzxml2search program of Trans-Proteomic Pipeline (Institute of Systems Biology, Seattle, WA) (Keller et al. 2005). In database search, common contaminant proteins were added to protein

databases, and reversed versions of all sequences were appended as decoy sequences for FDR estimation. The database search results were processed by IDPicker software for peptide validation and protein assembly. Throughout this study, IDPicker was configured to derive score thresholds to yield a 2% FDR. Peptides passing these thresholds were considered as legitimate identifications. Spectra for which these peptides were assigned were considered as "identified spectra". The datasets are available for download from Vanderbilt University Mass Spectrometry Research Center's web site (http://www.mc.vanderbilt.edu/msrc/bioinformatics/data.php).

"DLD1 LTQ" Dataset

This dataset was previously used to test IDPicker software and the experimental description was published by Ma et al. (2009). The "DLD1 LTQ" dataset consisted of four RPLC runs of human colon adenocarcinoma cells (DLD-1 cell line) analyzed on a Thermo Fisher LTQ linear ion trap mass spectrometer (San Jose, CA). The files averaged 12,913 MS/MS scans. Spectra were identified against an IPI human database (v3.56) using database search engines MyriMatch. Sequest and X!Tandem search results were converted to pepXML format using out2xml and tandem2xml programs in the Trans-Proteomic Pipeline, respectively. Raw peptide identifications were processed by the IDPicker software for protein assembly. Spectra were classified into two categories, "identified spectra" and "unidentified spectra", where the identified set pooled data from all three database searches.

Dataset name	# of files	(Average) # of MS/MS scans	Identification methods	Databases used for search
Removal of Low Quality Spectra				
DLD1 LTQ	4	12913	MyriMatch, Sequest, X!Tandem	IPI.HUMAN.v3.56
Mouse HCT	4	5408	MyriMatch, Sequest, X!Tandem	IPI.MOUSE.v3.62
Yeast Velos	5	38466	MyriMatch, Sequest, X!Tandem	SGD.orf_trans_all.20090303
Recovery of Unidentified High Quality Spectra				
DLD1 LTQ	1	12820	Sequest/MyriMatch, X!Tandem	IPI.HUMAN.v3.56
Serum Orbi	1	6697	MyriMatch, tryptic/semi-tryptic	IPI.HUMAN.v3.56
Histone Orbi	1	9170	MyriMatch/TagRecon	IPI.HUMAN.v3.68
Prediction of Richness of Identifiable Spectra				
MudPIT Orbi	10	9828	MyriMatch	IPI.HUMAN.v3.56
IEF Orbi	10	10897	MyriMatch	IPI.HUMAN.v3.56
GelBand LTQ	10	9520	MyriMatch	IPI.HUMAN.v3.47
Use of Quality Score in Peptide Validation				
DLD1 LTQ	4	12913	Mascot, Sequest, X!Tandem	IPI.HUMAN.v3.56
Selection of Spectra for De Novo Sequencing				
Yeast Velos	1	38560	PepNovo, MyriMatch	SGD.orf_trans_all.20090303
Tardigrade QSTAR	1	837	PepNovo, MyriMatch	SwissProt.DROME.ANOGA.CAEEL.rel56.8
Hadrosaur Orbi	1	14217	PepNovo, MyriMatch	AnoCar1.0
Use of ScanRanker in Cross-linking Analysis				
Crosslink Orbi	1	1161	Protein Prospector	SwissProt.ECOLI.20100810

Table 3. Experimental datasets for the evaluation of ScanRanker.

"Serum Orbi" Dataset

This dataset was previously used to test IDPicker software and the experimental description was published by Ma et al. (2009). The "Serum Orbi" data represented an RPLC analysis of depleted human serum sample in an LTQ-Orbitrap hybrid mass spectrometer (Thermo, Scan Jose, CA) at Vanderbilt University Medical Center. Spectra were identified against an IPI Human database (v3.56) using MyriMatch in either tryptic

or semi-tryptic search mode. Search results were processed by IDPicker and spectra were separated to three categories: “spectra identified in tryptic search”, “new identifications in semi-tryptic search” and “unidentified spectra”.

“Histone Orbi” Dataset

This dataset was published by Loecken et al. (2009). Histone H2b and H3 adducts was analyzed using an LTQ-Orbitrap mass spectrometer. Spectra were searched using MyriMatch against an IPI human database (v3.68) and processed by IDPicker for peptide validation and protein assembly. The identified proteins were pulled to construct a subset protein database for bind modification search by TagRecon.

“MudPIT Orbi” Dataset

This dataset was published by Slebos et al. (2008). Tryptic peptides from 50 µg proteins (adenocarcinoma) were loaded to a SCX column followed by a reverse phase LC-MS/MS analyses. Spectra from 10 fractions in the MudPIT experiment were searched using MyriMatch against an IPI Human database (v3.56) and processed by IDPicker.

“IEF Orbi” Dataset

This dataset was published by Slebos et al. (2008). Tryptic peptides from 50 µg proteins (adenocarcinoma) were separated by isoelectric focusing, followed by a reverse phase LC-MS/MS analyses. Spectra from 10 fractions in the IEF experiment were searched using MyriMatch against an IPI Human database (v3.56) and processed by IDPicker.

"GelBand LTQ" Dataset

This dataset was published by Burgess et al. (2008). Serum samples were collected from patients without evidence of malignancy. Alpha 2 macroglobulin-containing protein complexes were immunoprecipitated and separated by molecular weight in 10% SDS-PAGE. Each lane was sliced into 10 regions and subjected to in-gel digestion. Peptides from each gel region of each patient were subjected to a 95 minute RPLC separation. As peptides eluted in nanospray, the ions were directed to the inlet of a Thermo LTQ tandem mass spectrometer. Spectra were searched using MyriMatch against an IPI Human database (v3.47) and processed by IDPicker.

"Tardigrade QSTAR" Dataset

Hypsibius dujardini, a species of Tardigrades (commonly known as 'water bears') were grown in glass Petri dishes feeding on algae. Proteins from 600 organisms were collected and solubilized in LDS buffer (1M DTT), boiled, sonicated and then separated by 1D SDS-PAGE. Contiguous gel bands were excised, digested (trypsin), and samples were analyzed by reverse-phase nano-HPLC-ESI-MS/MS using an Eksigent nano-LC 2D HPLC system (Eksigent, Dublin, CA) which was directly connected to a quadrupole time-of-flight (QqTOF) QSTAR Elite mass spectrometer (MDS SCIEX, Concorde, CAN). Briefly, peptide mixtures were loaded onto a guard column (C18 Acclaim PepMap100, 300 µm I.D. x 5 mm, 5 µm particle size, 100 Å pore size, Dionex, Sunnyvale, CA) and washed with the loading solvent (0.1 % formic acid, flow rate: 20 µL/min) for 5 min. Subsequently, samples were transferred onto the analytical C18-nanocapillary HPLC column (C18 Acclaim PepMap100, 300 µm I.D. x 15 cm, 3 µm

particle size, 100 Å pore size, Dionex, Sunnyvale, CA) and eluted at a flow rate of 300 nL/min using the following gradient: 2-40% solvent B in A (from 0-35 min), 40-80% solvent B in A (from 35-45 min) and at 80% solvent B in A (from 45-55 min), with a total runtime of 85 min (including mobile phase equilibration). Solvents were prepared as follows, mobile phase A: 2% acetonitrile / 98% of 0.1% formic acid (v/v) in water, and mobile phase B: 98% acetonitrile / 2% of 0.1% formic acid (v/v) in water. Mass spectra (ESI-MS) and tandem mass spectra (ESI-MS/MS) were recorded in positive-ion mode with a resolution of 12000-15000 full-width half-maximum. For collision induced dissociation tandem mass spectrometry (CID-MS/MS), the mass window for precursor ion selection of the quadrupole mass analyzer was set to ± 1 *m/z*. The precursor ions were fragmented in a collision cell using nitrogen as the collision gas. Advanced information dependent acquisition (IDA) was used for MS/MS collection, including QSTAR Elite (Analyst QS 2.0) specific features, such as “Smart Collision” and “Smart Exit” (fragment intensity multiplier set to 4.0 and maximum accumulation time at 2.5 sec) to obtain MS/MS spectra for the three most abundant parent ions following each survey scan. Dynamic exclusion features were based on value M not *m/z* and were set to exclusion mass width 50 mDa and exclusion duration of 120 sec. Since complete genomic sequences for tardigrade are not yet available, I searched the dataset using MyriMatch against a database consisting of proteins from three taxonomically related species with complete proteomes, *Drosophila melanogaster* (DROME), *Anopheles gambiae* (African malaria mosquito, ANOGA) and *Caenorhabditis elegans* (CAEEL), downloaded from Swiss-Prot (release 56.8). Reversed sequences of these proteins were appended to the

database as decoys. Spectra were separately processed by PepNovo for *de novo* sequencing and ScanRanker for spectral quality assessment.

"Hadrosaur Orbi" Dataset

This "Hadrosaur Orbi" dataset represented an RPLC run of protein extracts from an 80-million-year-old Campanian hadrosaur, *Brachylophosaurus canadensis*, in a Thermo Fisher LTQ Orbitrap XL mass spectrometer published by Asara et al. (Schweitzer et al. 2009). The mzData file was downloaded from PRIDE (http://www.ebi.ac.uk/pride/, accession number 9285) and was converted to mzXML format using a predecessor of the MSConvert tool from the ProteoWizard library, which was subsequently processed by PepNovo and ScanRanker. Spectra were searched using MyriMatch against a lizard (*Anolis carolinensis*) database, AnoCar1.0, produced by the Broad Institute at MIT and Harvard (http://www.broadinstitute.org/models/anole). Common contaminant proteins were added to supplement these sequences, and reversed versions of all sequences were appended to complete the FASTA.

"Crosslink Orbi" Dataset

This dataset was provided by Robert Chalkley at University of California, San Francisco and published by Trnka, M. J. et al. (Trnka & Burlingame 2010). Purified GroEL and GroES proteins were cross-linked by 1,3-diformyl-5-ethynylbenzene (DEB). The sample was analyzed on an ESI LTQ-OrbitrapXL with an ETD module installed (Thermo Scientific). Cross-linked spectra were identified using Protein Prospector and were manually confirmed by Trnka et al. The dataset was also searched using Protein

Prospector against SwissProt *E.coli* database to identify spectra of non-crosslinked peptides. The search were performed with both parent and product mass tolerance of 20 ppm. Carbamidomethylcysteine was searched as a fixed modification. Methionine oxidation, protein N-terminal acetylation and peptide N-terminal glutamine cyclization to pyroglutamate were specified as variable modification.

"Mouse HCT" Dataset

This dataset was generated from a whole mouse liver protein extract obtained from adult CD1 mice in Vanderbilt University Mass Spectrometry Research Center. Proteins were reduced with DTE and alkylated with iodoacetamide prior to digestion with sequencing grade Trypsin. Four replicate LC-MS/MS runs were performed on a Bruker Esquire HCT ultra ion trap (Bruker Daltonics, Billerica, MA). The scan sequence consisted of 1 precursor ion scan (*m/z* = 375-1200) in standard enhanced and five subsequent tandem MS scans (*m/z* = 100-2800) in ultra scan mode. Scan averaging was set to 2 and ICC was 200,000. Singly charged peptides were excluded from tandem MS and dynamic exclusion was activated for 1 minute after two successful tandem MS experiments for a peptide. LC-MS/MS was carried out on an Agilent 1100 HPLC modified with a flow splitter and a FAMOS autosampler with a 2 µl sample loop. The column was a 12.5 cm, singly-vented, 360/75 um OD/ID PicoFrit emitter from New Objective attached to a 3 cm precolumn. Both columns were packed in house with 5 µm Monitor C18 particles. Each injection consisted of 6.5 ng of mouse liver digest. The mobile phases were water and acetonitrile with 0.1 % formic acid as an additive. Peptides eluted during the 60 minute gradient from 2 % to 50 % acetonitrile. Instrument raw data

were converted to mzXML format using the Bruker CompassXport tool. The files averaged 5048 MS/MS scans. Spectra were searched against an IPI mouse database (v3.62) by MyriMatch, Sequest and X!Tandem and processed as described in the "DLD1 LTQ" dataset.

"Yeast Velos" Dataset

This dataset was generated utilizing the CPTAC Yeast Performance Standard that was digested with trypsin in Rapigest (Paulovich et al. 2010). Two microliter portions of peptide mixture were analyzed using a Velos ion trap mass spectrometer (Thermo, San Jose, CA) equipped with an Eksigent 1D Plus NanoLC pump and Eksigent NanoLC-AS1 autosampler (Eksigent, Dublin, CA). Peptides were solid-phase extracted using an in-line column (100 μm × 6 cm) packed with Jupiter C18resin (5 μm, 300 Å, Phenomenex, Torrence, CA) and separated on a capillary tip (100 μm × 11 cm, Polymicro Technologies, Phoenix, AZ) packed with the C18 resin. Following the injection, peptides were solid-phase extracted by washing with 0.1% FA (mobile phase A) for 15 min at a flow rate of 1.5 μL/min. Mobile phase B consisted of acetonitrile (ACN) with 0.1% FA. Peptides were separated using a gradient of 2−40% B for 120 min at a flow rate of 700 nL/min, followed by a rapid increase of B from 40−90% in 25 min, and held at 90% B for 9 min before returning to initial conditions of 100% A. Survey scans were collected in the ion trap a mass range of 400−2000 *m/z*. Following each survey scan, the five most intense ions were selected for MS/MS fragmentation in the ion trap using the dynamic exclusion feature (exclusion mass width of -1 *m/z* and +2 *m/z*, exclusion duration of 60 s, and repeat count of 1). Centroided MS/MS scans were acquired on the Velos using an

isolation width of 2 *m/z*, an activation time of 30 ms, an activation *q* of 0.250 and a normalized collision energy of 30 using 1 microscan with a max ion time of 100 ms for each MS/MS scan and 1 microscan with a max ion time of 50 ms for each full MS scan and a minimum signal of 1000. The mass spectrometer was tuned prior to analysis using the synthetic peptide TpepK (AVAGKAGAR), and the tune parameters were as follows: spray voltage of 1.5 kV, a capillary temperature of 200 °C and an S-lens RF level of 59%. The MS/MS spectra were collected using data-dependent scanning in which one full MS spectrum was followed by four MS-MS spectra. MS/MS spectra were recorded using dynamic exclusion of previously analyzed precursors for 60 s with a repeat count of 1 and a repeat duration of 1. A total of five replicate LC-MS/MS experiments were performed and 192,330 MS/MS spectra were collected. Spectra were searched using MyriMatch, Sequest and X!Tandem against the *Saccharomyces* Genome Database orf_trans_all.fasta file downloaded in March of 2009 and processed by IDPicker.

III.4 Results and Discussion

To establish the effectiveness of ScanRanker in quality estimation, I first evaluated its three metrics for discrimination. After establishing its scoring discrimination, I tested its real-world performance for recognition of unidentified high quality spectra and prediction of richness of identifiable spectra. I also demonstrated its applications in peptide validation, *de novo* sequencing and cross-linking analysis. These tests establish ScanRanker as a robust and effective algorithm for spectral quality assessment of data from various instruments in a wide variety of applications.

III.4.1 Subscore Evaluation

ScanRanker evaluates spectral quality based on "BestTagScore", "BestTagTIC" and "TagMzRange" subscores. To test the effectiveness of subscores, the "DLD1 LTQ" (Ma et al. 2009) dataset was searched by MyriMatch, Sequest and X!Tandem to maximize the peptide identifications. The discriminating power of each subscore is illustrated via receiver operating characteristic (ROC) curves in Figure 15. Each subscore may be used to discriminate spectral quality between identified and unidentified spectra. By combining the three subscores, however, ScanRanker achieves better discrimination than by using any single subscore alone. Results obtained after testing any combination of two subscores were exceeded by combining all three subscores (data not shown).

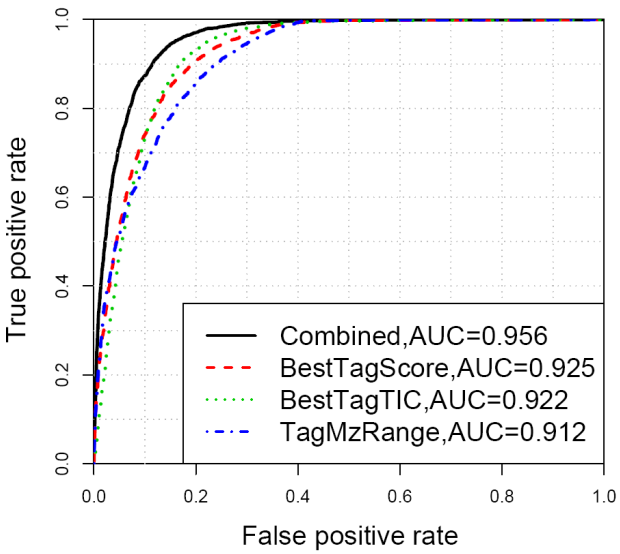

Figure 15. Combining three subscores improves the discriminating power of ScanRanker. Tests on the "DLD1 LTQ" dataset revealed different discrimination in ScanRanker's subscores. The ROC curves display true positive rate (a.k.a. sensitivity) and false positive rate (a.k.a. 1-specificity) of ScanRanker's subscores and the combined score. The AUC values show that combining three subscores yields better discrimination than using any single subscore.

I tested both mean and median for subscore normalization during the development of ScanRanker algorithm, and they worked equally well because of small differences between these values. For example, the average difference between mean and median of "DLD1 LTQ" dataset (4 replicates) are 1%, 6% and 2% for "BestTagScore", "BestTagTIC" and "TagMzRange" subscores, respectively. I chose the mean of subscores for normalization because it is less expensive to compute than the median. More importantly, if ScanRanker scores need to be adjusted across multiple files, the mean of subscores across these files can be easily calculated based on the sum of subscores and the total count of spectra.

ScanRanker computes the quality score by averaging three normalized subscores. If the subscores differed considerably in their discriminating powers, simply averaging the subscores would reduce the discriminating power of ScanRanker overall. To test the discrimination difference between optimized score weights and equal weights, each subscore was assigned a weight from 0 to 1 with 0.1 increments, and the summation of weighted subscores was used to calculate the area under ROC curve (AUC). The best possible weighting yielded an AUC less than 1% higher than the equal weight approach. As a result, I opted to use equal weights for simplicity.

III.4.2 Removal of Low Quality Spectra

Low quality spectra, particularly from ion trap mass spectrometers, often generate a significant amount of computational overhead but contribute little to protein identification. Filtering these spectra via ScanRanker prior to search can save time in identification. To test ScanRanker's performance in removing low quality spectra, I

analyzed three datasets collected from a Thermo Fisher LTQ, an Esquire HCT ultra and a Thermo Fisher LTQ Velos ion trap. MyriMatch searched these data in two ways: (1) search all spectra, (2) only search the top 60% of high quality spectra as reported by ScanRanker. In all three instruments, more than 94% of the resulting identifications were shared between both searches, and more spectra were identified in the second search than in the first. In the case of the Esquire HCT, almost 5% of the identifications were produced only when the bottom 40% of spectra were pruned away, at the cost of less than 1% of the identifications (see Figure 16). More identifications were gained by removing low quality spectra prior to database search; low quality spectra are more prone to be matched to decoy sequences, thus increasing the stringency of the threshold applied to all identifications.

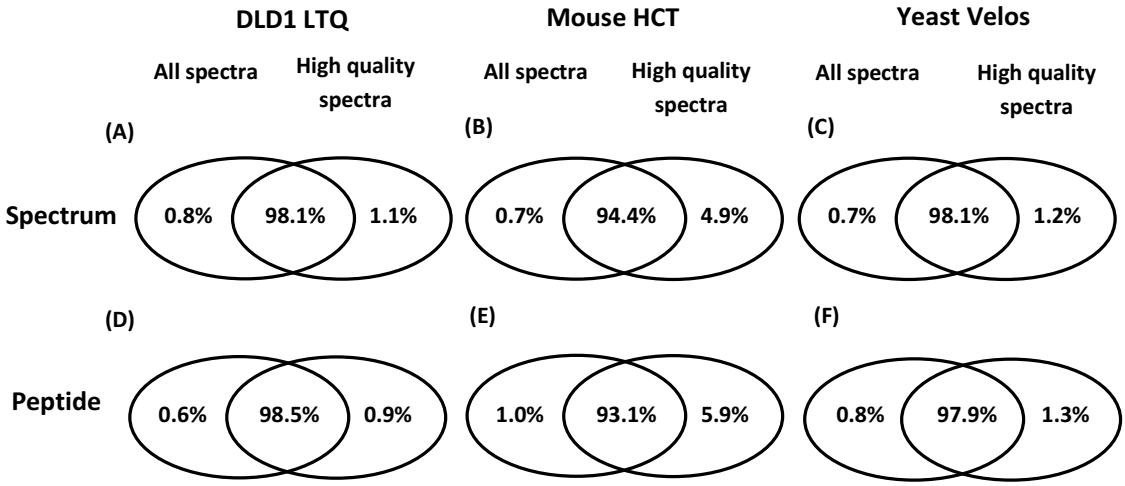

Figure 16. Removing poor MS/MS scans in ScanRanker does not significantly reduce identifications. Panels A-C show the percent overlap of identified spectra when searching either all spectra or only high quality spectra. Similar overlaps for identified peptides are displayed in Panels D-F.

Although I retained the top 60% spectra in our test, it should be noted that there is no common threshold that can be applied to all datasets for the selection of high quality spectra. The spectral removal will be more beneficial for large-scale proteomics studies in which multiple biological and technical replicates are analyzed. I recommend determining the percentage of retained spectra by examining the search results of all spectra from a single replicate, then applying the threshold to remove low quality spectra in other replicates. For example, Figure 17 plots the proportion of retained identified spectra in context of spectra sorted by ScanRanker scores. It is obvious that the top ranked 60% spectra in all three datasets contain more than 95% of identified spectra. Therefore, this threshold could be subsequently used to remove low quality spectra in other replicates before the database search. These figures can be easily generated from ScanRanker output, which comprises a tab-delimited text file including ranked spectra, identification labels and the cumulative sum of identification labels.

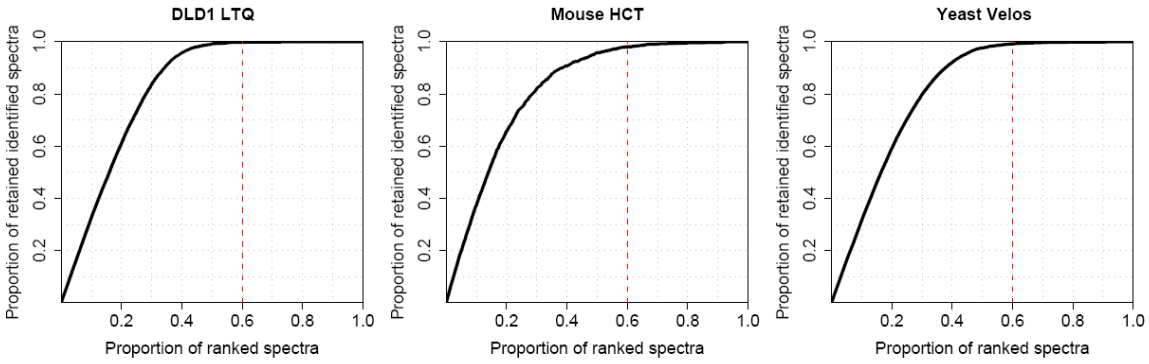

Figure 17. Determine spectral removal threshold from a single replicate.

III.4.3 Recovery of Unidentified High Quality Spectra

Simple database search can sometimes fail to identify many spectra that can be identified through additional effort. I employed three publicly available datasets to determine if ScanRanker scores were predictive of identifications gained through more advanced searching methods.

In the first test, I evaluated the peptides identified through multiple database search algorithms. A single replicate in the “DLD1 LTQ” dataset with 12820 MS/MS scans was analyzed using Sequest, yielding 2878 confidently identified spectra. Additional searches using MyriMatch and X!Tandem identified 826 new spectra missed in the Sequest search. All spectra were sorted by ScanRanker scores from high to low quality and were split into deciles. Figure 18A shows the number of initially identified spectra, newly identified spectra and unidentified spectra in each decile. As expected, identified spectra, either by Sequest or additional searches, were associated with higher ScanRanker scores than unidentified spectra.

The second experiment evaluated the peptides gained through semi-tryptic search. For samples dominated by a few major proteins, this strategy improves peptide and protein identification. In this study, I searched the “Serum Orbi” (Ma et al. 2009) dataset using MyriMatch in either fully tryptic or semi-tryptic search mode. Among 6697 MS/MS scans in the dataset, 646 spectra were identified in tryptic search, and an additional 928 spectra were generated by semi-tryptic search. Figure 18B plots the distribution of all spectra, split to deciles by ScanRanker scores. It can be observed that the majority of gained spectra by semi-tryptic search were ranked within the top 30% of spectra by ScanRanker.

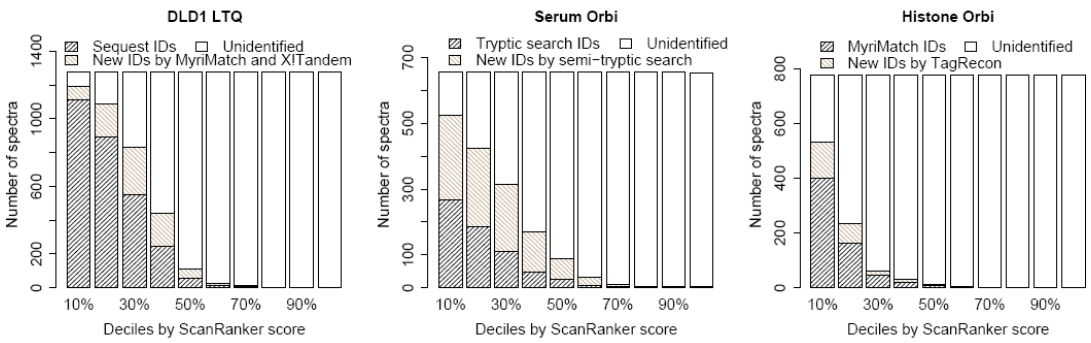

Figure 18. Evaluation of ScanRanker to recover unidentified high quality spectra. Three datasets were reanalyzed by additional search methods to find high quality spectra that were unidentified in initial database searches. Each test represents a typical reason that high quality spectra may be left unidentified in an initial search. (A) The "DLD1 LTQ" dataset was initially identified by Sequest search. New identifications (IDs) were added by MyriMatch and X!Tandem searches. (B) The "Serum Orbi" data was searched by MyriMatch in either tryptic or semi-tryptic mode. (C) The "Histone Orbi" data was searched by MyriMatch. A subsequent TagRecon search was performed to identify spectra of mutated or modified peptides. These graphs plot the distributions of initial identifications, new identifications by additional searches and unidentified spectra in deciles by ScanRanker scores. In each panel, the left side represents spectra assigned high ScanRanker quality scores and the right side is low quality spectra. Newly identified spectra tend to associate with better ScanRanker scores in all datasets.

In the third test, I examined the ability of ScanRanker to find spectra that were unidentified due to modifications and mutations. The "Histone Orbi" (Loecken et al. 2009) data with 9170 MS/MS scans was initially searched using MyriMatch, yielding 641 confidently identified spectra. To find spectra of modified peptides, the dataset was searched using TagRecon against a customized database consisting of identified proteins and decoy sequences. TagRecon yielded 672 spectra including common modifications such as acetylation (117 spectra) and deamidation (159 spectra). Among them, 234 spectra were missed in MyriMatch search. Figure 18C shows the distribution of spectra ordered by ScanRanker scores. As in preceding plots, spectra assigned high ScanRanker scores were more likely to be identified through PTM identification software.

III.4.4 Comparison of ScanRanker to QualScore

QualScore is a tool integrated in the Trans-Proteomic Pipeline that is specifically designed for recognizing spectra that evade identification. I compared the performance of QualScore and ScanRanker on three datasets. To obtain quality scores from QualScore, I analyzed the datasets using Sequest and PeptideProphet, and then processed results using QualScore under the default configuration. Figure 19 shows the ROC curves of ScanRanker and QualScore in three datasets. ScanRanker performed as reliably as QualScore in all tests. ScanRanker displayed slightly better performance than QualScore in the "Histone Orbi" data, possibly because the existence of modified peptides decreased the effectiveness of Sequest/PeptideProphet training, thus diminishing QualScore accuracy. Despite this minor difference, both tools are able to recognize unassigned high quality spectra. QualScore produces accurate results by training its scoring system for each dataset based on Sequest/PeptideProphet results, while ScanRanker evaluates spectral quality directly using a sequence tagging approach. Thus, ScanRanker has no dependence on the availability of database search results.

I attempted to include other algorithms in this comparison. Initial tests of msmsEval gave promising discrimination for LTQ datasets, but no training model was provided to enable its use in other types of instruments. The version of the PARC filter (Bern et al. 2004) that I received from the Yates Laboratory omitted scores for removed spectra, limiting its scope to filtering spectra prior to database search. In some other tools, the software simply split datasets to "good" and "bad" directories without a report of metrics for each spectrum, limiting conclusions about their scoring discrimination. As a result of these setbacks, I limited the comparison to QualScore.

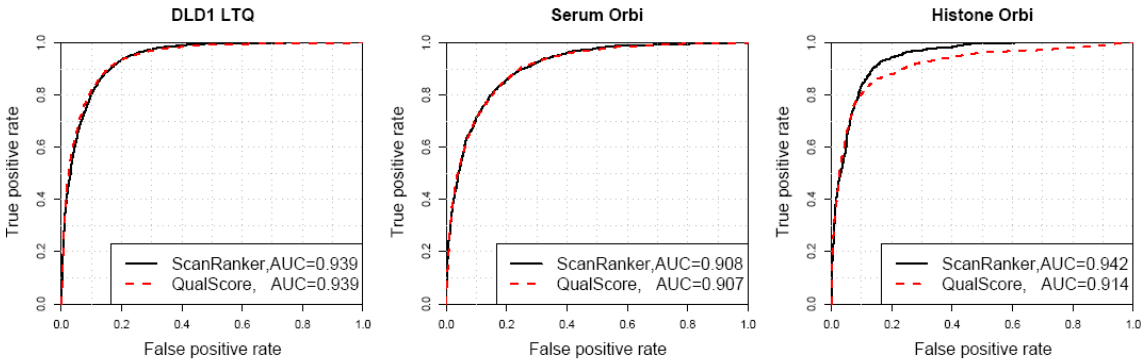

Figure 19. Comparison of ScanRanker to QualScore. Spectra in three datasets were separately processed by ScanRanker and QualScore to generate quality scores. ScanRanker performs as well as QualScore in all datasets but does not require Sequest/PeptideProphet analysis for spectral quality assessment.

III.4.5 Prediction of Richness of Identifiable Spectra

High quality spectra are more likely to be identified in proteomics data analysis. If multiple LC-MS/MS runs are included in an experiment, (for example, MudPIT or 1D gel experiments,) the number of high quality spectra in each dataset reveals the richness of identifiable spectra, providing a preliminary overview for the quality of the LC-MS/MS experiment. I sought to demonstrate that the ScanRanker scores are predictive of relative qualities of LC-MS/MS runs in an experiment. Three published datasets, the "MudPIT Orbi" (Slebos et al. 2008), "IEF Orbi" (Slebos et al. 2008) and "GelBand LTQ" (Burgess et al. 2008) data, were searched using MyriMatch against an IPI Human database. ScanRanker grouped all LC-MS/MS runs in each dataset as a single experiment, in which the means and interquartile ranges of subscores across all fractions or gel bands were used for normalization to compute the quality scores.

Figure 20 shows the scatter plot between the number of identified spectra in each LC-MS/MS run and the number of retained spectra with ScanRanker scores above different thresholds. Here I used three score thresholds (0, 0.5 and 1). Spectra with score 0 represent scans of better than 60-70% spectra, and spectra scoring 0.5 and 1 have better quality than approximately 85% and 95% of spectra in each experiment, respectively. The distributions of quality scores, however, are dataset-dependent. As expected, the number of high quality spectra predicted by ScanRanker in each dataset is highly correlated to the number of identified spectra. For example, a score threshold at 0.5 produced the Pearson correlation coefficients of 0.90, 0.90 and 0.95 for “MudPIT Orbi”, “IEF Orbi” and “GelBand LTQ” datasets, respectively. Therefore, the relative quality of each LC-MS/MS run in an experiment can be estimated by the number of high quality spectra determined by ScanRanker. This is potentially useful for large-scale proteomic studies, in which ScanRanker can be used as a rapid quality control tool to highlight bad LC-MS/MS runs among an experiment.

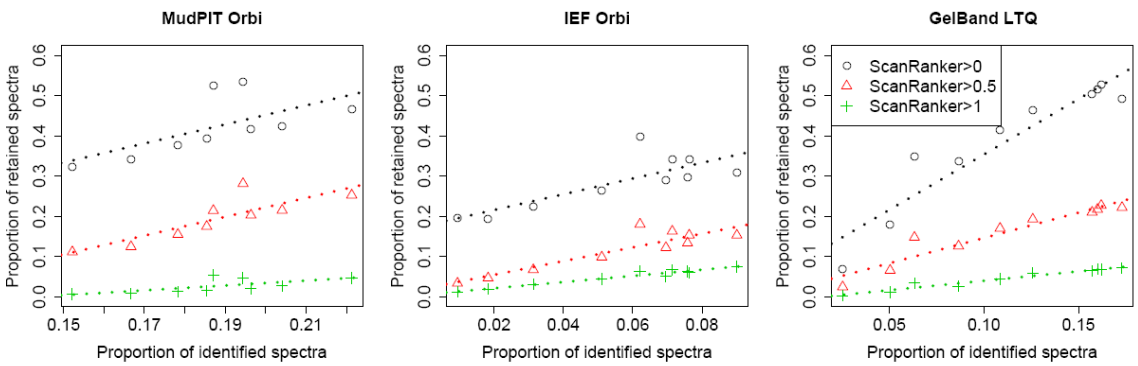

Figure 20. ScanRanker scores predict the richness of identifiable spectra. Each point in the figure represents a single LC-MS/MS run and the dotted lines show the least squares fit of the data. Three ScanRanker thresholds were used to count retained spectra. 9 of 10 LC-MS/MS runs in the MudPIT dataset are plotted because the first fraction of the MudPIT experiment generated only 21 spectrum identifications. Each LC-MS/MS run in all three datasets includes about 10000 MS/MS spectra, while the number of identified spectra varies dramatically. The number of spectra assigned high ScanRanker scores correlate to the number of identified spectra, providing relative quality assessment of LC-MS/MS runs in an experiment.

III.4.6 Use of Quality Score in Peptide Validation

In proteomics data analysis, database search engines usually generate one or more scores to measure the matches between candidate peptides and experimental spectra. The search results are then processed by either statistical methods (e.g., PeptideProphet) or FDR-based methods (e.g., IDPicker) for peptide validation. In latter methods, usually only scores from database search tools are used to compute FDR. Here I sought to combine spectral quality scores and scores produced by database search tools to increase confident peptide identifications. I searched the "DLD1 LTQ" data using Mascot, Sequest and X!Tandem against an IPI Human database (v3.56). All search results were converted to pepXML files using either an in-house Perl script or software tools in the Trans Proteomics Pipeline. The spectral quality scores generated by ScanRanker were added to pepXML files using a Perl script. IDPicker subsequently read these scores along

with search engine scores during peptide validation. The software combined multiple scores by optimizing score weights through a Monte Carlo method, generating a single score for each peptide-spectrum match. In this test, I configured IDPicker to use either the primary scores from a database search tool or these scores plus the spectral quality score.

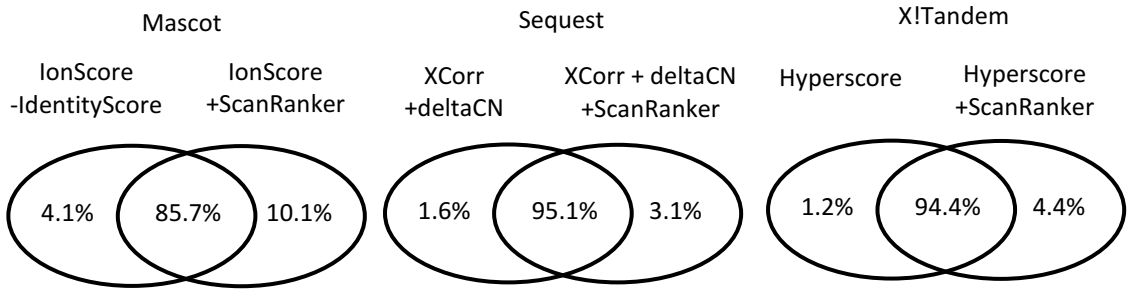

Figure 21. Adding ScanRanker scores in peptide validation increases the number of confident spectrum identifications. “DLD1 LTQ” dataset was separately searched by Mascot, Sequest and X!Tandem. ScanRanker scores were added to pepXML files to allow score combination in IDPicker. Mascot scores were combined using either static weights as “IonScore-IdentityScore” or optimized weights as “IonScore + ScanRanker”. Sequest and X!Tandem results were combined by enabling score weights optimization in IDPicker. The Venn diagrams show the percent overlap of identified spectra when using either a single score or combination of two scores. The latter method yielded more spectrum identifications for all searches.

Figure 21 shows the percent overlap of confident spectrum identifications in both settings. Adding spectral quality scores in peptide validation consistently yielded more confident spectrum identifications than using a single score. Mascot benefited significantly more from score combination than Sequest and X!Tandem. Some spectra may be identified only when using the primary score. These spectra, however, are usually less confident identifications that are assigned marginal match scores in database search.

III.4.7 Selection of Spectra for *De Novo* Sequencing

De novo sequencing is an alternative, database-independent approach for peptide identification. However, inferring peptides from spectra is a time-consuming process. In this study, for example, PepNovo took about 8 hours to infer sequences of an Orbitrap dataset with 14217 MS/MS scans on a Dell Optiplex 745 computer with an Intel Core 2 Duo 6400 processor and 3 GB of RAM, while ScanRanker only required 3 minutes for spectral quality assessment. Therefore, *de novo* sequencing could benefit from the application of spectral quality assessment tools by selecting high quality spectra for *de novo* analysis.

As a state-of-the-art *de novo* sequencing tool, PepNovo assigns a score to each inferred peptide sequence to evaluate how well it explains the peak pattern in a spectrum. The higher a PepNovo score, the better an inferred peptide matches a spectrum. I employed three datasets to demonstrate that high ScanRanker scores are predictive of high PepNovo scores. The initial comparison of these scores analyzed the “Yeast Velos” dataset, in which peptide identification was straightforward. Figure 22A shows the scatter plot between the PepNovo score of the top ranked peptide sequence for each spectrum and its ScanRanker score. ScanRanker scores are highly correlated to PepNovo scores, producing a Pearson correlation coefficient of 0.82. As expected, spectra identified by MyriMatch search tend to associate with high ScanRanker and PepNovo scores.

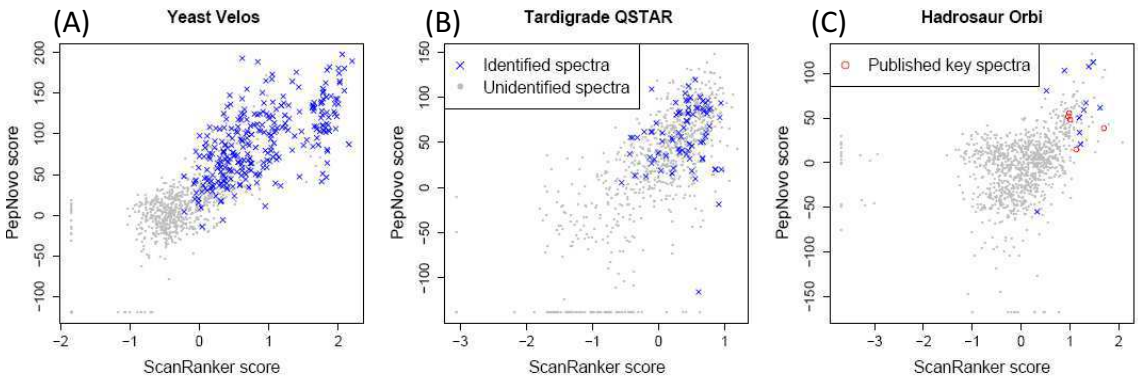

Figure 22. ScanRanker scores can be used to predict *de novo* sequencing success. Spectra in three datasets were separately processed by ScanRanker and PepNovo. Identifications were generated by searching the spectra using MyriMatch. For clarity, only 1000 spectra were randomly sampled and displayed. When PepNovo reported no peptide for a spectrum, it was visualized as matching the minimum score reported by the software for that dataset. Panel C highlights five published key spectra from the Asara group publication. In all three tests, spectra with high ScanRanker scores tend to be assigned high PepNovo scores, implying that ScanRanker can be used to select high quality spectra for *de novo* sequencing.

Next, I evaluated ScanRanker on datasets for which *de novo* sequencing would be necessary. The "Tardigrade QSTAR" dataset is an LC-MS/MS experiment from a 1D gel band from a species of microscopic animals for which genome sequence is unavailable. MyriMatch attempted to produce identifications in a customized database containing proteins of three species that are taxonomically similar to tardigrade (*Drosophila melanogaster* (DROME), *Anopheles gambiae* (African malaria mosquito, ANOGA) and *Caenorhabditis elegans* (CAEEL)). Only spectra for peptides of highly similar proteins would be identified by this approach; only 66 spectra were identified among the 837 MS/MS scans in the set. Figure 22B superimposes these identifications on the scatter plot of PepNovo and ScanRanker scores. PepNovo and ScanRanker both report that many spectra were of high quality and yet failed identification. Pearson correlation between the two algorithms produced a coefficient of 0.72.

Considerable controversy has accompanied the recent publication of proteomics data for fossilized specimens (Schweitzer et al. 2009). I sought to characterize the recent "Hadrosaur Orbi" dataset to evaluate the inherent identifiability of spectra for these spectra. I began with a database search against a lizard (*Anolis carolinensis*) database, AnoCar1.0, produced by the Broad Institute (http://www.broadinstitute.org/models/anole). The result included 189 confidently identified tandem mass spectra, but all matched to keratin or trypsin sequences (our database did not include the chicken sequences employed by the Asara group). I plotted spectra against the corresponding PepNovo and ScanRanker scores (see Figure 22C). Five collagen spectra from the original Asara publication were assigned high ScanRanker quality scores of 1.13, 0.99, 0.97, 1.01 and 1.70; I was unable to match the sixth identification to the corresponding MS/MS spectrum. The hadrosaur data produced the lowest correlation between PepNovo and ScanRanker (0.34), where the best correspondence could be observed in the high scoring domains for the two algorithms. It becomes clear that the data of the "Hadrosaur Orbi" set were disproportionately likely to produce PepNovo scores below zero, suggesting that a large fraction of spectra from this dataset could not support confident sequence identifications even if appropriate sequences were available in FASTA.

III.4.8 Use of ScanRanker in Cross-linking Analysis

Identification of cross-linked peptides by mass spectrometry is a challenging task, mainly because of the high complexity and often low signal intensity in these spectra. Even with the availability of advanced computational tools, manual interpretation or confirmation of cross-linked peptides is generally necessary. Here I sought to

demonstrate that ScanRanker helps to prioritize spectra for manual inspection. The published “Crosslink Orbi” (Trnka & Burlingame 2010) dataset consists of 1161 MS/MS spectra collected on an LTQ-Orbitrap XL with an ETD module installed (Thermo Scientific). Spectra in quadruply charged or higher charge states were selected for ETD fragmentation to characterize chemically cross-linked GroEL-GroES chaperonin complex. Protein Prospector (Chu et al. 2010) identified 55 spectra of cross-linked peptides (manually confirmed) and 91 spectra of single peptides. Figure 23 shows the distribution of these spectra, split to deciles by ScanRanker scores. The spectra of cross-linked peptides were associated with high ScanRanker scores, suggesting that ScanRanker is capable of recognizing these spectra, though they are more complicated than spectra of single peptides. The results also indicate that ScanRanker performs well for spectra from ETD fragmentation.

Some spectra were assigned high quality scores but remained unidentified. A manual inspection of these spectra implies that they are likely produced by peptides rather than non-peptide contaminants. These spectra usually contain a large number of peaks. For example, the top 10% of spectra by ScanRanker includes 70 unidentified spectra. The average number of peaks in these spectra is 228, which is much higher than that number of all spectra (91 peaks) in the dataset.

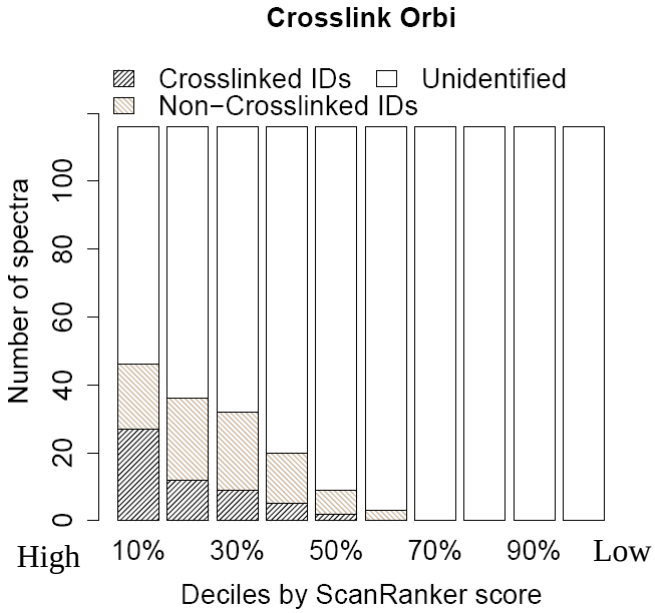

Figure 23. ScanRanker helps to prioritize spectra for manual inspection in cross-linking analysis. The "Crosslink Orbi" dataset was processed using Protein Prospector to identify crosslinked and non-crosslinked spectra. The figure plots the distribution of these spectra in deciles by ScanRanker scores. The identified spectra, either crosslinked or non-crosslinked, were associated with high ScanRanker scores, implying that ScanRanker can be used to facilitate cross-linking analysis by ranking spectra for manual inspection.

III.5 Conclusion

I present a method that assesses quality of tandem mass spectra through sequence tagging. ScanRanker does not require training for each type of data from different mass spectrometers, broadening its use to lab researchers lacking prior experience in statistical learning. In this study, I employed a variety of datasets to demonstrate the effectiveness of ScanRanker for recovery of unidentified high quality spectra and removal of low quality spectra. I showed that ScanRanker can be used to predict the richness of identifiable spectra in LC-MS/MS experiments and to improve peptide validation. I also demonstrate the application of our method to rank spectra for *de novo* sequencing and

cross-linking analysis. The superior performance of ScanRanker established it as a robust and reliable spectral quality assessment tool.

Wrapping ScanRanker to a library function will improve its usability, making it easy to be integrated into other software tools. For example, it can be used as a pre-processor for database search engines to filter out low quality spectra; it can be integrated to IDPicker to provide quality scores for spectral identifications and export unidentified spectra for subsequent analysis; it can be incorporated into QuaMeter (described in next chapter) to replace the identification step and conduct instrument QC based on the identifiable spectra rather than identified spectra.

CHAPTER IV

QUAMETER: MULTI-VENDOR PERFORMANCE METRICS FOR LC-MS/MS PROTEOMICS INSTRUMENTATION

IV.1 Introduction

Technologies for proteomic identification via LC-MS/MS rely on a complex series of experiments: protein denaturation and digestion, LC separation of peptides followed by electrospray ionization, tandem mass spectrometry, and proteome informatics. Variation in the performance for any of these elements may impact proteomic identification. The publication of LC-MS/MS quality metrics by Paul Rudnick at NIST, working in collaboration with the National Cancer Institute (NCI) CPTAC network, introduced a set of metrics that span this complex process (Rudnick et al. 2010), enabling recognition of components that were operating at variance with their typical performance. The strategy makes use of defined quality control samples that are periodically analyzed between experimental samples in a queue for the mass spectrometer.

The previously described 46 metrics embodied in the NIST MSQC software rely on a complex set of algorithms. Data from Thermo RAW files are first transcoded to mzXML, MS1, and MGF formats for subsequent processing. The MS1 files enable peptide precursor ion chromatograms to be assessed in the NIST ProMS software. The tandem mass spectra of an LC-MS/MS experiment are identified by either the SpectraST spectral library search engine or the OMSSA data-base search algorithm. The MSQC

software can then match precursor ion chromatograms with peptide identifications to compute its set of metrics and report them to a text file.

In practice, several aspects of the MSQC software prevent its use for routine instrument monitoring. Its reliance on a modified ReAdW tool for reading raw data limits its application to instruments from Thermo Fisher. The coordination among different software packages may lead to mis-association of peptide identifications and tandem mass spectra when alternative file formats or high scan rate instrumentation are employed. Adapting the pipeline for site-specific workflows (such as a different peptide identification engine) is a non-trivial task.

In this work, I present the QuaMeter tool that has the same capabilities as MSQC with several important additions. QuaMeter can read files from most mass spectrometry vendors via ProteoWizard and does not lose time trancoding to other formats. The software accepts identification data from IDPicker, so any identification database search engine that produces pepXML or mzIdentML can be used. I demonstrate the use of QuaMeter for data collected from instrument of three different vendors. I examine the impact of identifications tools on computed metrics. The improvements in QuaMeter make it a robust and flexible quality metric assessor with open source.

IV.2 Overview

To compute QC metrics for a LC-MS/MS experiment, QuaMeter requires two input files: an instrument spectral file and an identification search engine results file. As shown in Figure 24, ProteoWizard is central to data management for the software pipeline. Its support for native file formats from multiple instrument vendors means that

transcoding data to an open format is unnecessary (although support for non-Microsoft Windows systems would require this step). ProteoWizard presents spectra in an mzML data model to all of the pictured tools, using uniform "nativeID" labels to relate identifications to source scans. Its incorporation of a chromatogram extractor from Crawdad (Finney et al. 2008) supports the full-width-at-half-maximum computations needed for the quality metrics.

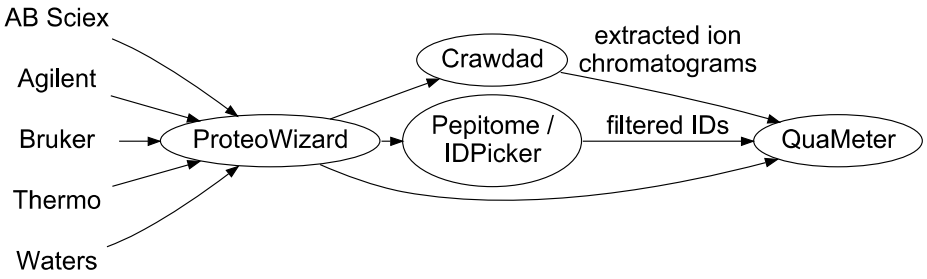

Figure 24. Workflow diagram for QuaMeter operation.

The peptide identification tools such as MyriMatch (database search), TagRecon (sequence tag-based database search), and Pepitome (spectral library search) all incorporate ProteoWizard for both data import and identification export via pepXML and mzIdentML formats. Here I emphasize Pepitome since spectral library search is particularly appropriate for repeat identification of QC standard samples. Raw identifications from this step are filtered within the IDPicker protein assembler, and filtered identifications are processed by QuaMeter to compute QC metrics. For each LC-MS/MS run, QuaMeter exports metrics to a table in text format.

IV.3 Data Sources

I tested QuaMeter on several datasets spanning six instruments from three different vendors (see Table 4). These datasets were accumulated via QC experiments in three laboratories to monitor instrumentation performance..

Dataset	Sample	# of files	Average # of MS/MS scans	Databases used for search
LTQ-XL	BSA	280	11917	RefSeq.BOVINE
LTQ-Orbitrap	BSA	53	3417	RefSeq.BOVINE
LTQ-Velos	Yeast	5	38466	SGD.orf_trans_all
HCT Ultra	BSA	24	3467	RefSeq.BOVINE
QSTAR Elite	Beta-gal	23	451	UniProt.ECOLI
TripleTOF 5600	Beta-gal	60	1973	UniProt.ECOLI

Table 4. Experimental datasets for the evaluation of QuaMeter.

All datasets were searched using MyriMatch or Pepitome, and search results were processed by IDPicker software for peptide validation and protein assembly. Throughout this study, IDPicker was configured to derive score thresholds to yield a 5% FDR. Filtered identifications and spectral files were processed by QuaMeter to compute QC metrics. To compare metrics generated by QuaMeter and MSQC, scripts in Awk were created to make IDPicker identifications accessible to MSQC so that both algorithms could work from a common set of identifications. Data processing details and software parameters are available in Appendix A.

Thermo Fisher LTQ-XL Dataset

This data constitutes of 280 routine BSA runs at the Jim Ayers Institute for Precancer Detection and Diagnosis at Vanderbilt University. The dataset was previously

used to test the Pepitome software and the experimental description is published by Dasari et al. (Dasari et al. 2012). The files average 11917 MS/MS scans each. All files were searched using MyriMatch against a RefSeq BOVINE database or using Pepitome o match the NIST BSA spectral library (http://peptide.nist.gov).

Thermo Fisher LTQ-Orbitrap Dataset

This dataset was also collected at the Jim Ayers Institute for Precancer Detection and Diagnosis at Vanderbilt University. Experimental settings were exactly the same as above except 10x BSA peptide mixtures were used instead of 1x BSA. All samples were analyzed on a Thermo Fisher LTQ-Orbitrap mass spectrometer. A total of 53 files were used in this manuscript. Spectra were searched using MyriMatch against a RefSeq BOVINE database. The files average 3417 MS/MS scans.

Thermo Fisher LTQ-Velos Dataset

This is the same dataset as described above for testing the ScanRanker software. Five technical replicates were collected for a yeast lysate on a Thermo Fisher LTQ-Velos instrument. The files average 38466 MS/MS scans each. Spectra were identified using MyriMatch against a yeast database (http://www.yeastgenome.org) downloaded on March 2009. All files were also searched by Pepitome against the NIST yeast spectral library (http://peptide.nist.gov).

Bruker Daltonics HCT Ultra Dataset

Stock BSA solution prepared in 100mM ammonium bicarbonate buffer was digested overnight with sequencing grade Trypsin (Promega) at enzyme-to-substrate ratio of 1:50 at 37°C. LC-MS/MS analysis was carried out on an Eksigent 1D-nanopump coupling to a Bruker HCT Ultra iontrap mass spectrometer. The mobile phases were water and acetonitrile with 0.1% formic acid as an additive. 2uL of working BSA solution of 100fmol/uL was load by a FAMOS autosampler with a 10uL sample loop onto a 3cm, 360/100 OD/ID trap column of 5um Jupiter C18 particles with loading aqueous buffer of 0.1% formic acid at flow rate of 1uL/min and separated on a 15cm 360/75um OD/ID PicoFrit emitter column from New Objective packed with 3um Jupiter C18 particles. Both columns were in house packed. Tryptic peptides eluted during a gradient from 2% to 50% acetonitrile at flow rate of 250nL/min. Different LC-gradients were applied throughout the data collection. LC-MS/MS data was acquired in positive ionization mode with scan segments of 1 precursor ion scan (*m/z*=375-2000) in standard enhanced and 3 subsequent tandem MS scans of three most abundant ions in ultra scan mode. Scan average was set to 2 and ion charge control (ICC) was 200,000. Singly charge ions were excluded from tandem MS and a 1 minute dynamic exclusion was activated for each peptide after two MS tandem acquisitions.

Instrument raw files were converted to mzML format by the MSConvert tool in ProteoWizard. Since Bruker data extraction library does not write precursor spectrum reference information in mzML files, which is required for running QuaMeter, a Perl script is created to add precursor spectrum references to MS/MS scans. The latest previous MS1 scan is assumed as the precursor of neighboring MS/MS scans. 24 files

were collected with averagely 3467 MS/MS scan each. All spectra were searched using MyriMatch against a RefSeq BOVINE protein database and identifications were filtered by IDPicker.

AB SCIEX QSTAR Elite Dataset

Predigested, tryptic beta-galactosidase solutions (*E. coli*) were obtained from AB SCIEX and used as quality control samples. Samples were analyzed by reverse-phase nano-HPLC-ESI-MS/MS using an Eksigent nano-LC 2D HPLC system (Eksigent, Dublin, CA) which was directly connected to a quadrupole time-of-flight (QqTOF) QSTAR Elite mass spectrometer (AB SCIEX, Concord, CAN). Briefly, peptide mixtures were loaded from the autosampler (using partial loop fill methods) onto a guard column (C18 Acclaim PepMap100, 300 µm I.D. x 5 mm, 5 µm particle size, 100 Å pore size, Dionex, Sunnyvale, CA) and washed with the loading solvent (0.1 % formic acid, flow rate: 20 µL/min) for 5 min. Subsequently, samples were transferred onto the analytical C18-nanocapillary HPLC column (C18 Acclaim PepMap100, 75 µm I.D. x 15 cm, 3 µm particle size, 100 Å pore size, Dionex, Sunnyvale, CA) and eluted at a flow rate of 300 nL/min using the following gradient: 2-30% solvent B in A (from 0-15 min), 30-80% solvent B in A (from 15-17 min) and at 80% solvent B in A (from 17-20 min), with a total runtime of 52 min (including mobile phase equilibration). Solvents were prepared as described below for the TripleTOF 5600. Mass spectra (ESI-MS) and tandem mass spectra (ESI-MS/MS) were recorded in positive-ion mode with a resolution of 12,000-15,000 full-width half-maximum. For collision induced dissociation tandem mass spectrometry (CID-MS/MS), the mass window for precursor ion selection of the

quadrupole mass analyzer was set to ± 1 *m/z*. The precursor ions were fragmented in a collision cell using nitrogen as the collision gas. Advanced information dependent acquisition (IDA) was used for MS/MS collection, including QSTAR Elite (Analyst QS 2.0) specific features, such as "Smart Collision" and "Smart Exit" (fragment intensity multiplier set to 2.0 and maximum accumulation time at 2.5 sec) to obtain MS/MS spectra for up to seven most abundant precursor ions following each survey scan. Dynamic exclusion features were based on value M not *m/z* and were set to exclusion mass width 50 mDa and exclusion duration of 60 sec. All 23 files were searched using MyriMatch against a UniProt *E.coli* database and identifications passing 5% FDR in IDPicker analysis were confident IDs.

AB SCIEX TripleTOF 5600 Dataset

Predigested, tryptic beta-galactosidase solutions (*E. coli*) were obtained from AB SCIEX and used as quality control samples. Samples were analyzed by reverse-phase HPLC-ESI-MS/MS using an Eksigent Ultra Plus nano-LC 2D HPLC system (Dublin, CA) which was directly connected to a new generation quadrupole time-of-flight (QqTOF) TripleTOF 5600 mass spectrometer (AB SCIEX, Concord, CAN) in direct injection mode. The autosampler was operated in full injection mode overfilling a 1 µl loop with 3 µl analyte for optimal sample delivery reproducibility. Briefly, after injection, peptide mixtures were transferred onto the analytical C18-nanocapillary HPLC column (C18 Acclaim PepMap100, 75 µm I.D. x 15 cm, 3 µm particle size, 100 Å pore size, Dionex, Sunnyvale, CA) and eluted at a flow rate of 300 nL/min using the following gradient: at 5% solvent B in A (from 0-13 min), 5-35% solvent B in A (from 13-29 min), 35-80%

solvent B in A (from 29-31 min) and at 80% solvent B in A (from 31-37 min), with a total runtime of 58 min including mobile phase equilibration. Solvents were prepared as follows, mobile phase A: 2% acetonitrile/98% of 0.1% formic acid (v/v) in water, and mobile phase B: 98% acetonitrile/2% of 0.1% formic acid (v/v) in water. Mass spectra and tandem mass spectra were recorded in positive-ion and "high-sensitivity" mode, with a resolution of ~35,000 full-width half-maximum in MS1 mode and ~15,000 in MS/MS mode. The nanospray needle voltage was 2,400 V in HPLC-MS mode. After acquisition of ~ 5 to 6 samples, TOF MS spectra and TOF MS/MS spectra were automatically calibrated during dynamic LC-MS & MS/MS autocalibration acquisitions injecting 25 fmol beta-galactosidase. For collision induced dissociation tandem mass spectrometry (CID-MS/MS), the mass window for precursor ion selection of the quadrupole mass analyzer was set to ± 1 *m/z*. The precursor ions were fragmented in a collision cell using nitrogen as the collision gas. Advanced information dependent acquisition (IDA) was used for MS/MS collection on the TripleTOF 5600 (Analyst TF 1.5) to obtain MS/MS spectra for the 20 most abundant precursor ions following each survey MS1 scan (allowing for 50 msec acquisition time per each MS/MS). Dynamic exclusion features were based on value M not *m/z* and were set to an exclusion mass width of 50 mDa and an exclusion duration of 15 sec. All 60 files were searched using MyriMatch against a UniProt *E.coli* database and processed by IDPicker.

IV.4 Results and Discussion

IV.4.1 Differences between QuaMeter and MSQC

Validating QuaMeter performance began with a comparison of the values computed by MSQC and QuaMeter. I modified MSQC to accept the same identified peptides from IDPicker as did QuaMeter. BSA QC runs collected on a Thermo Fisher LTQ-XL mass spectrometer were identified by Pepitome using the NIST ion trap spectral library (http://peptide.nist.gov), and IDPicker filtered the results to a 5% FDR. Scripts converted the filtered identifications for MSQC handling.

Figure 25 illustrates the correspondence between QuaMeter and MSQC outputs for a set of representative metrics. Median precursor *m/z* error for +2 peptides (MS1-5A in NIST nomenclature) is shown in the top-left panel as a representative of metrics with very good agreement between both implementations. Most metrics representing peptide identifications (such as P-2A, P-2B, P-2C and P-3) yielded similar results.

The key C-2A metric was a note of discord between QuaMeter and MSQC. This metric, describing the duration of time in which the middle 50% of peptides are identified, disagreed even when QuaMeter attempted to emulate MSQC behavior closely (top-right panel in Figure 25). Inspection of the code revealed that MSQC vacillates in whether or not modifications or precursor charge differentiate identifications. Because C-2A plays a role in the computation of many other metrics, the QuaMeter implementation was changed to a “distinct modified peptide” rule (under which either a sequence difference or a modification change resulted in the identification counting as a new peptide). Since distinct modified forms for a peptide sequence may chromatographically elute differently, this change leads to a more representative metric.

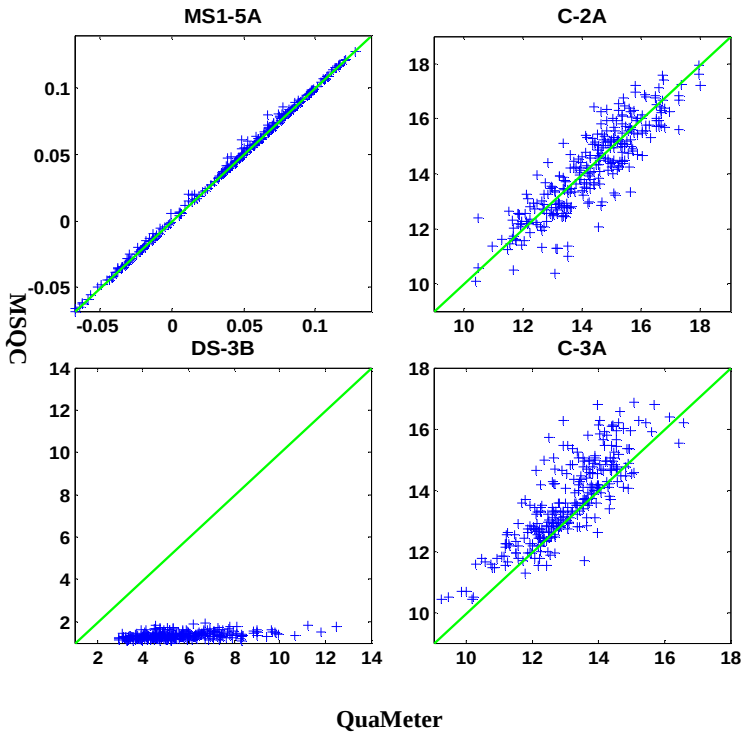

Figure 25. QuaMeter generates similar metrics as MSQC except several chromatographic metrics due to the use of distinct chromatogram extraction tools. Metrics were generated from BSA QC experiments collected on a Thermo Fisher LTQ-XL mass spectrometer.

Because MSQC and QuaMeter extract chromatographic data by distinct tools, differences in peak intensity and width are unsurprising. Metric DS-3B evaluates the maximum intensity versus the intensity at the time when MS/MS was triggered for the 50% of peptides with the least intense trigger intensities (see bottom-left panel in Figure 25). The MSQC software estimated far lower peak intensity maxima than expected from manual inspection, resulting in little correlation for this metric. This effect propagated through metrics describing the chromatographic process as well as dynamic sampling. Metric C-3A (lower-right panel in Figure 25) reports the median peak width (FWHM) for identified peptides. QuaMeter, via Crawdad, generally reports lower peak widths than

does the MSQC code. It should be noted that this comparison was performed using an early version of MSQC that uses a modified ReAdW tool for chromatogram extraction. This strategy has been deprecated in updated MSQC in favor of the ProMS tool that may produce more reliable chromatographic data (P. Rudnick, personal communications). I was unable to acquire a recent build of ProMS for comparative testing.

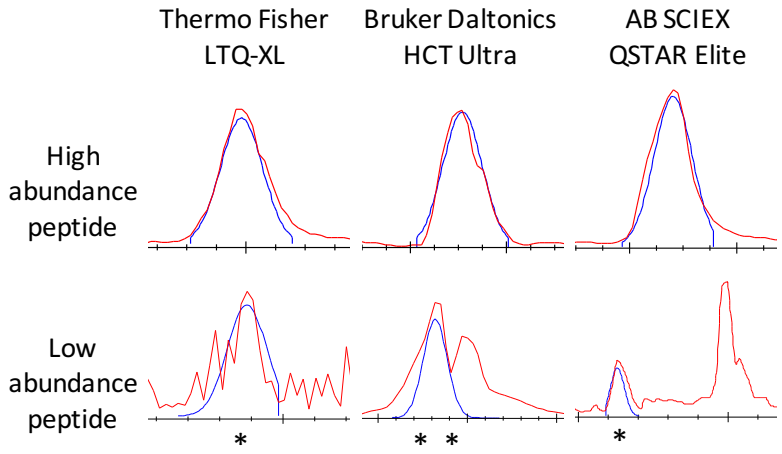

Figure 26. QuaMeter generates reliable chromatographic data in instruments from multiple vendors via the Crawdad function in ProteoWizard. Red lines represent experimentally measured intensities in MS and blue lines are extracted ion chromatograms generated by Crawdad. Asterisks for the low abundance peptides signify the acquisition times for identified MS/MS scans.

Because the Crawdad function has been implemented in the ProteoWizard library, QuaMeter can extract chromatographic data from all major vendor formats. QuaMeter provides an option to export chromatographic data in mz5 format (Wilhelm et al. 2012) which can be visualized by the SeeMS tool in ProteoWizard. Figure 26 illustrates the extracted ion chromatograms (XIC) of experimentally measured intensities and modeled peaks generated by Crawdad. XIC of representative peptides from three instrument

platforms were displayed. For high abundance peptides that were identified with many MS/MS scans, Crawdad produced well-fitted chromatograms that match experimental data (top panels in Figure 26). In addition, Crawdad also showed excellent performance for low abundance peptides with noisy experimental XIC or interfering peaks (bottom panels in Figure 26). QuaMeter chromatogram extraction is improved by using the precursor mass calculated from identified peptides and by noting the retention times of identified MS/MS scans.

IV.4.2 Multi-vendor Performance

To test QuaMeter's compatibility with instruments from multiple vendors, I employed several datasets collected from Thermo Fisher LTQ-XL, LTQ-Orbitrap, LTQ-Velos, Bruker Daltonics HCT Ultra, AB SCIEX QSTAR Elite and AB SCIEX TripleTOF 5600 mass spectrometers. Instrument raw files from Thermo and Bruker were converted to mzML format using the MSConvert tool in ProteoWizard. AB SCIEX data were converted to mzML files using the AB SCIEX MS Converter (version 1.2). All data were searched by MyriMatch and search results were processed by IDPicker. Filtered identifications were then processed by QuaMeter to compute QC metrics.

Rather than evaluate instrumentation performance solely based on the number of identifications, QuaMeter provides six categories of QC metrics that monitor chromatographic performance, electrospray source stability, MS1 and MS2 signals, dynamic sampling of ions and peptide identification. For example, Figure 27 illustrates a set of selected metrics describing the chromatographic process for five instruments.

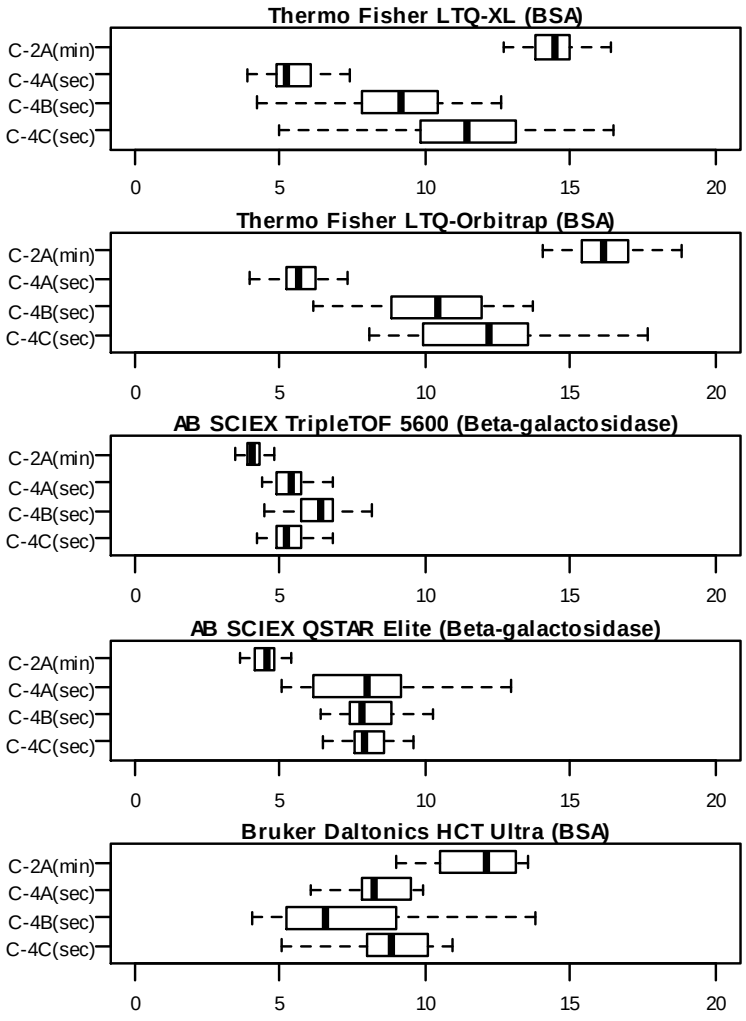

Figure 27. QuaMeter computes QC metrics for multiple instrument platforms. Standard samples such as BSA or beta-galactosidase were analyzed for routine instrument evaluation. C-2A: time period over which middle 50% of peptides were identified. C-4A, C-4B, C-4C: median peak width for identified peptides in first, last and median RT decile.

These plots reflect experimental settings and reveal instrument performance variability. First, the C-2A metric, the duration of time in which the middle 50% of peptides are identified, is very small for the AB SCIEX TripleTOF 5600 and QSTAR Elite dataset, implying that peptides were eluted in a short time period. This is because a very short LC gradient was applied for peptide separation in these experiments. Second, the variation of C-2A metric is relatively large for the Bruker Daltonics HCT Ultra data.

This is because different BSA samples for this instrument were separated by different HPLC columns and gradients. Likewise, large variations were also observed for other QC metrics computed for this dataset such as the number of identifications (data not shown). Third, peak widths of identified peptides were not evenly distributed across retention time in all tests. The C-4A, C-4B and C-4C metrics report median peak width for identified peptides in the early, late and middle retention time, respectively. It can be observed that peak width for all instruments varies with retention time. These plots demonstrate the cross-instrument capabilities of ProteoWizard and QuaMeter.

QuaMeter metrics can be used to spot abnormal instrument performance. For example, early analysis of TripleTOF 5600 data recognized six files as outliers compared to other QC experiments. As shown in Figure 28, very low numbers of identifications were generated from these six files (top-left panel), and a close examination of QuaMeter metrics showed that they associated with high precursor mass accuracy errors (bottom-left panel). The instrument log revealed that these files had a mass accuracy shift due to temperature variation (caused by air handler failure within the laboratory). Recalibrating these files yielded narrow precursor errors and comparable number of identifications as other experiments (right panels in Figure 28).

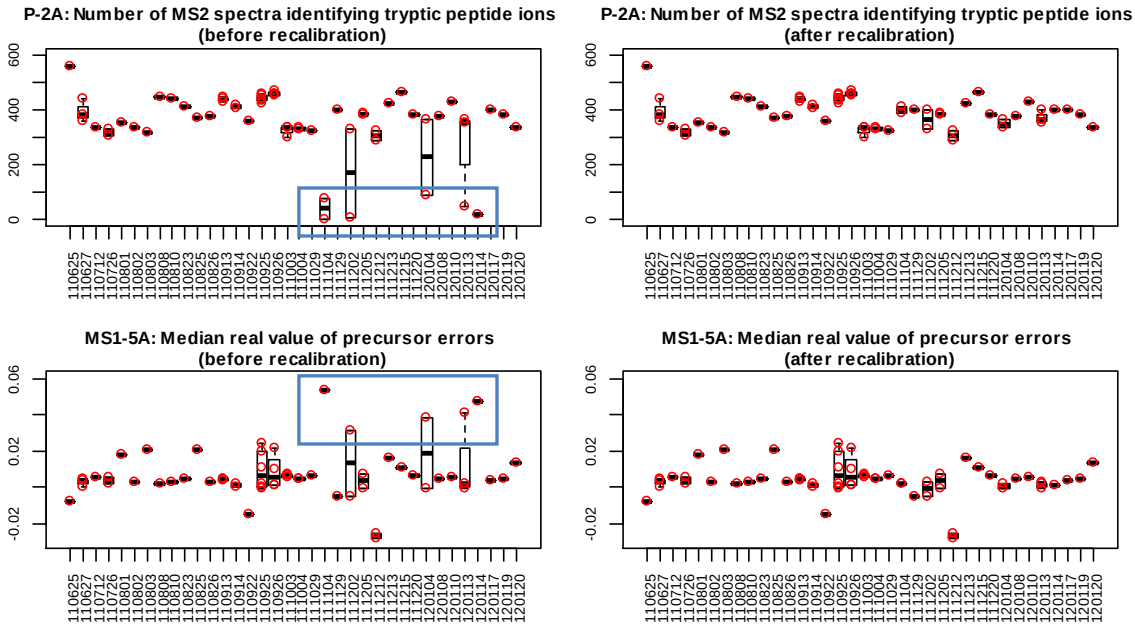

Figure 28. QuaMeter metrics help to spot abnormal instrument performance. Metrics computed from TripleTOF data were plotted by date. Six files were recognized as outliers in early analysis that had very low number of identifications (blue box in top-left P-2A metric) and high precursor mass accuracy errors (blue box in bottom-left MS1-5A metric; one point missing for 111104 because zero identification passed 5% FDR threshold from this file). Recalibrating these files yielded narrow precursor errors (bottom-right panel) and comparable number of identifications as other experiments (top-right panel).

IV.4.3 Impact of identification tools

Because QuaMeter relies on identified peptides to compute QC metrics, different tools for identification may yield different QuaMeter metrics. To evaluate this impact in generating QC metrics, I employed a yeast lysate dataset with five technical replicates analyzed on a Thermo Fisher LTQ-Velos mass spectrometer. The files average 38466 MS/MS scans each. This test demonstrates that QuaMeter works well not only for simple samples such as BSA and beta-galactosidase but also for complex mixtures. Spectra were identified either through database search by MyriMatch or through spectral library search by Pepitome. Both identification tools exported search results in pepXML format for

processing by IDPicker for peptide validation and protein assembly. Filtered identifications were then read by QuaMeter for QC evaluation. Because it accepts filtered identifications from IDPicker, any workflow in which identification tools produce search results in pepXML or mzIdentML format can also support QC.

Figure 29 plots a set of selected QuaMeter metrics, one from each of the six categories, computed based on MyriMatch and Pepitome identifications. Some metrics shifted when the source of identifications changed. For example, the spectral library search by Pepitome identified around 15% more spectra than using MyriMatch (see P-2A in Figure 29). However, changing identification tools does not lead to substantial changes for most metrics. In addition, although identification tools produced different QC metrics, the variation for the five replicates from MyriMatch search resembled that seen from Pepitome. Therefore, it is very likely that the identification tool has limited effect in accessing analytical system performance and technical variability. Given the fact that a spectral library search is usually much faster than a typical database search, Pepitome, coupled with QuaMeter, provides a practical solution for routine identification and analysis of standard QC samples.

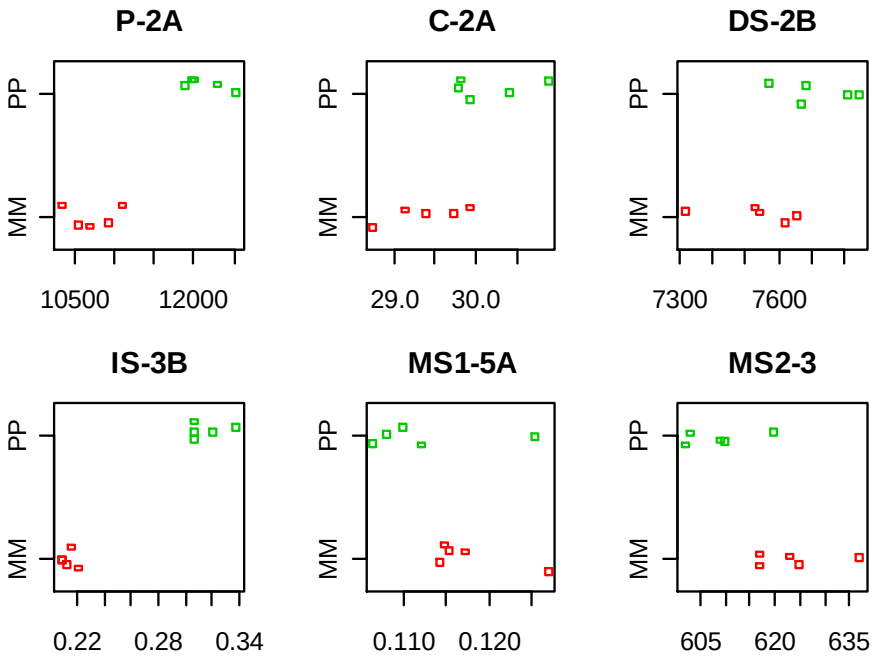

Figure 29. Distinct identification tools produce different QC metrics with similar variation. Five technical replicates of yeast lysate samples were analyzed on a Thermo Fisher LTQ-Velos mass spectrometer. Spectra were identified by MyriMatch (MM) and Pepitome (PP) separately. Identifications from each search engine were used to compute QC metrics. P-2A: Number of MS2 spectra identifying tryptic peptide ions; C-2A: Time period over which middle 50% of peptides were identified; DS-2B: Number of MS2 scans taken over C-2A. IS-3B: Number of 3+ peptides over 2+ peptides; MS1-5A: Median real value of precursor errors; MS2-3: Median number of peaks in MS2 scans.

IV.5 Conclusion

I presented an open-source tool that computes objective metrics for the evaluation of shotgun proteomics instrumentation performance. QuaMeter advances the previous MSQC tool by supporting most mass spectrometer vendors via the use of the ProteoWizard library. The ability to work with IDPicker identification data allows it to be incorporated to any identification workflow that produces pepXML or mzIdentML files. The improvements in QuaMeter make it a reliable and flexible tool for shotgun proteomics QC analysis.

Although QuaMeter supports native file formats from multiple instrument vendors, many time native files are converted to open formats in data analysis pipeline. QuaMeter requires the same spectral file that the database search engine was fed with. In additional, drawing conclusions from QuaMeter output is less well-established now. Another ongoing project is developing a statistical method based on QuaMeter metrics to enable on-the-fly instrument QC. A subset of key metrics need be determined to evaluate the analytical systems in routine practice.

Future directions for QuaMeter include a number of goals. First, recording metrics for experiments to a database rather than a collection of text files will greatly improve the production utility of the software. Second, incorporating assessments of MS/MS quality by ScanRanker would be much faster and more adaptable than incorporating peptide identifications. Optimizing the strategies by which metric values can be evaluated to diagnose sources of instrument variability will be essential. As these techniques mature, QC metrics promise to automate recognition of instrument inconsistency before critical samples are wasted.

CHAPTER V

DISCUSSION

V.1 Summary of Results

The work in this dissertation described three new software tools for shotgun proteomics data analysis (see Figure 30). The QuaMeter tool focuses on instrumentation quality control to assure that data fed into analysis pipeline are collected under stable instrument performance. The IDBoost tool focuses on rescuing spectral identifications after initial data analysis. Spectra that are not identified after IDBoost can be further recovered by the ScanRanker tool for advanced searches. Each tool was developed to solve one aspect of problems, but together they work coordinately to provide an improved shotgun proteomics data analysis pipeline.

The IDBoost tool provides a simple and efficient way to rescue spectral identifications from current analysis. It incorporates identification evidence of similar spectra and applies a rating method to determine the majority vote of these spectra. Spectra that were discarded in original analysis due to the failure of passing confidence threshold can be rescued in subsequent data analysis. Meanwhile, IDBoost corrects database search errors by taking into account search results from a cluster of similar spectra. In this dissertation, I demonstrated its applications in phosphorylation studies and spectral count based comparative analysis. IDBoost helps to solve phosphorylation site ambiguity and improves spectral count based quantification. In addition, IDBoost was

implemented in IDPicker, which provides a graphical user interface for interactive validation of rescued identifications.

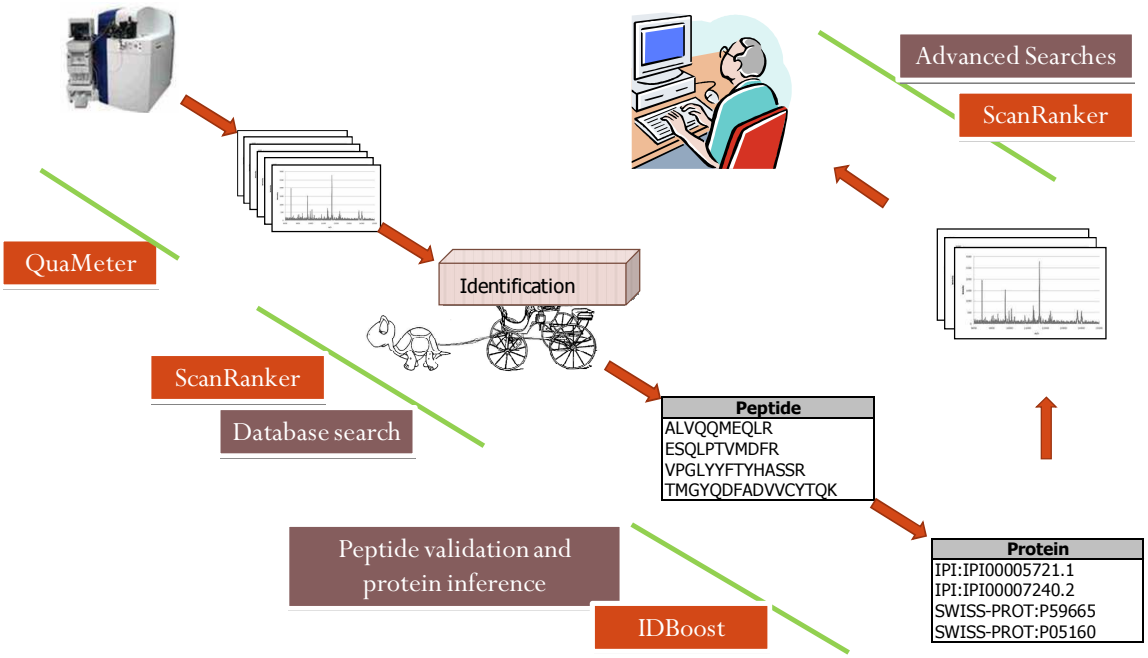

Figure 30. A summary of three bioinformatics tools in proteomics data analysis workflow.

IDBoost expands spectral identifications based on existing analysis. This is one of its advantages because it does not require additional identification steps. However, this also limits its usage because no new peptides will be added to the analysis. In practice, IDBoost should be used as a quick tool to rescue identifications after IDPicker analysis. To find more peptides, a subset of unidentified spectra can be exported by ScanRanker for subsequent advanced searches.

The ScanRanker tool is a tandem mass spectral quality assessor that recognizes the potentially identifiable MS/MS scans. The core of ScanRanker is the DirecTag sequence tagging program. ScanRanker evaluates the quality of tandem mass spectra by

examining how well sequence tags can be inferred from each spectrum. It can be used to recognize unidentified high quality spectra after IDPicker and IDBoost analysis. In this dissertation, I demonstrated the use of ScanRanker to select spectra for *de novo* sequencing and cross-linking analysis. This tool can also be used to predict the richness of identifiable spectra among multiple LC-MS/MS runs in an experiment. ScanRanker is particularly useful for analyzing samples lacking accurate genome annotations. To improve its usability, I made a GUI to run ScanRanker and also a program to view the results. The IonMatcher viewer allows interactive validation of peptide-spectrum-matches and offers an interface for *de novo* sequencing of unidentified spectra.

The QuaMeter tool is another quality assessor but focuses on the evaluation of analytical systems rather than tandem mass spectra. The goal of this project is to provide a quality control tool that can be used in many labs for routine instrument monitoring. QuaMeter supports most mass spectrometry vendors via ProteoWizard and does not require transcoding instrument raw files to other format, making it a fast and easy-to-use QC assessor. Because it works with IDPicker identifications, QuaMeter is flexible to be integrated into any existing workflow that generates pepXML or mzIdentML. In this dissertation, I demonstrated the use of QuaMeter for data collected from different vendors.

V.2 Future Direction

V.2.1 Peptide Identification

A critical component in proteome informatics is the scoring system that interprets observed spectra to peptide sequences. New technologies emerged in the past few years

have greatly increased the quality of proteomics data, while the informatics tools is slowly catching up. For example, ETD and CID are complementary fragmentation methods to cover both long and short peptides. Peptide ions can now be selectively fragmented on either ETD or CID mode (Swaney et al. 2008). However, it is challenging to combine the information obtained by ETD and CID in a single search. Because they generate totally different fragment ions, different scoring schemes need to be developed to process ETD and CID data separately. Even the same scoring method can be adapted to work for both fragmentation methods, the score distribution may differ substantially, making it difficult to combine the search results.

Current scoring systems usually only consider major fragment ions such as b and y ions for CID fragmentation. Adding other abundant ions in scoring schemes may improve the discrimination power between correct and incorrect assignments. For example, LTQ-Orbitrap data under HCD fragmentation retains low mass and immonium ions, which can be considered in scoring methods to improve peptide identification.

Understanding gas-phase fragmentation chemistry is very important for the development of scoring schemes. Unfortunately, current fragmentation model in most database search tools predict theoretical spectra far away from the experimental data. The massive amount of MS/MS data being confidently interpreted and deposited to public repositories helps the development of sophisticate fragmentation models to predict accurate fragment ions and their intensities.

V.2.2 PTM Identification and Validation

Identification of PTM is still a challenging issue even with the recent improvements in MS instrumentation and enrichment methods. For example, one problem to identify phosphopeptides is that a large number of phosphopeptides undergoes beta-elimination reaction (loss of phosphate group) rather than backbone fragmentation under CID. This reduces fragment ion signals in MS/MS spectra, making it difficult to identify these peptides. As a result, complementary techniques such as ETD or MS/MS/MS spectrum may be required for data acquisition. In recent years, experimental platforms have been greatly improved for PTM analysis, while no substantial progress has been made in computational tools. Advanced algorithms that take advantage of the state-of-art technologies are desirable for accurate and large-scale PTM analysis.

PTM validation is also a difficult problem. Conventional validation methods may not be appropriate for PTM validation because the assumptions for these methods are likely to be violated. For example, the target-decoy based FDR approach assumes a one-to-one correspondence between incorrect target hits and decoy hits, while there is a much lower prior likelihood of observing a modified peptide compared to a non-modified peptide. As a result, the error rate estimated for the analysis may not be accurate. Future developments on advanced methods for PTM validation are necessary.

V.2.3 Next Generation Sequencing and Proteomics

Interpretation of proteomics data relies heavily on the protein databases, which are usually translated from genome DNA sequences. Over the past few years, there have been remarkable advances in DNA sequencing technologies with the rapid evolution of

next-generation sequencing (NGS). The advent of NGS has significantly increased the throughput and reduced sequencing cost by orders of magnitude, making it a cost-effective option to obtain global genomic information of the same biological system that is targeted for proteomics experiments.

The availability of complete genomics sequences from a species or individual facilitates MS-based protein identifications. The DNA sequences from the same system can be translated to proteins, resulting in a more accurate protein database for proteomics data analysis. An alternative way is to obtain the transcriptome (RNA-Seq) data, which may be a better representation for proteins with mutations or splice variants. With the cost reduction of NGS in the next few years, it is possible to routinely sequence critical samples and use customized protein databases for proteomics data analysis.

V.2.4 Integration of Omics Data

Proteomics alone may not be sufficient to characterize the complexity of biological systems. Recent advances in various omics technologies enable the detection of various biological molecules in a high-throughput manner. Combining different omics results obtained from the same biological system will substantially increase the understanding of complex biological process. Such a success has been demonstrated in the field of microbiology (Zhang et al. 2010), plant systems biology (Fukushima et al. 2009) and mouse organ protein profiling (Kislinger et al. 2006).

The Cancer Genome Altas (TCGA) initiative has a rich collection of human cancer genome data. Recently the NCI CPTAC consortium partnered with TCGA to integrate proteomics and genomics data for cancer research. The same tumor specimens

studied by the TCGA network will be analyzed by the CPTAC network, generating a pair of proteomic and genomic data for each sample. These different types of systematic measurements offer insights in how specific gene alterations affect proteins in individual tumors. Computational tools to integrate different omics data will play a critical role in these studies.

V.2.5 Targeted Proteomics

While whole proteome analyses have considerable appeal in systems biology, it has some practical limits such as relatively low dynamic range. Targeted proteomics, especially multiple reaction monitoring (MRM), are emerging to be a promising approach that provides greater dynamic range and higher confidence in identifications, which is particularly useful for biomarker verification. MRM methods are under active development in recent year, requiring the continuous development of bioinformatics tools. Algorithms for peptide and transition selection may benefit from mining the vast amount of identifications in the spectral libraries. Methods to detect quantification errors and estimate the experiment error rate are also desirable.

APPENDIX A

SOFTWARE CONFIGURATIONS

MyriMatch Configurations

Thermo Fisher LTQ-XL and LTQ-Velos, Bruker Daltonics HCT Ultra data:

PrecursorMzTolerance= 1.25
PrecursorMzToleranceUnits = daltons
FragmentMzTolerance = 0.5
FragmentMzToleranceUnits = daltons
AdjustPrecursorMass = false
NumSearchBestAdjustments = 3
DuplicateSpectra = true
UseChargeStateFromMS = false
NumChargeStates = 3
UseSmartPlusThreeModel = false
TicCutoffPercentage = 0.95
CleavageRules = "trypsin"
NumMaxMissedCleavages = 2
NumMinTerminiCleavages = 2
UseAvgMassOfSequences = true
MinCandidateLength = 5
DynamicMods = "M ^ 15.9949 (Q * -17.026" (add [STY] $ 79.9663 for phosphopeptide search)
MaxDynamicMods = 2
StaticMods = "C 57.0215"
ComputeXCorr = true

Thermo Fisher LTQ-Orbitrap data:

PrecursorMzTolerance= 10
PrecursorMzToleranceUnits = ppm
FragmentMzTolerance = 0.5
FragmentMzToleranceUnits = daltons
AdjustPrecursorMass = true
MinPrecursorAdjustment = -1.008665
MaxPrecursorAdjustment = 1.008665
PrecursorAdjustmentStep = 1.008665
NumSearchBestAdjustments = 3
DuplicateSpectra = true

UseChargeStateFromMS = true
NumChargeStates = 4
UseSmartPlusThreeModel = false
TicCutoffPercentage = 0.95
CleavageRules = "trypsin"
NumMaxMissedCleavages = 2
NumMinTerminiCleavages = 2
UseAvgMassOfSequences = false
MinCandidateLength = 5
DynamicMods = "M ^ 15.9949 (Q * -17.026" (add [STY] $ 79.9663 for phosphopeptide search)
MaxDynamicMods = 2
StaticMods = "C 57.0215"
ComputeXCorr = true

AB SCIEX data:

PrecursorMzToleranceRule = “mono”
AvgPrecursorMzTolerance = 1.5 *m/z*
MonoPrecursorMzTolerance = 100 ppm for QSTAR Elite and 50 ppm for TripleTOF
MonoisotopeAdjustmentSet = [-1,2]
FragmentMzTolerance = 0.4 *m/z* for QSTAR Elite and 0.05 *m/z* for TripleTOF
StaticMods = "C 57.0215"
DynamicMods = "M ^ 15.9949 (Q * -17.026"
MinTerminiCleavages = 1
CleavageRules = "Trypsin/P"
MaxMissedCleavages = 2
MaxDynamicMods = 2
DecoyPrefix = "rev_"
NumChargeStates = 3
OutputFormat= "pepXML"
SpectrumListFilters = "peakPicking false 2-"
TicCutoffPercentage = 0.98
FragmentationAutoRule = true
MaxResultRank = 5
MinPeptideMass = 0 Da
MaxPeptideMass = 10000 Da
MinPeptideLength = 5
MaxPeptideLength = 75
UseSmartPlusThreeModel = false
ProteinSampleSize = 100
ComputeXCorr = true
UseMultipleProcessors = true

Sequest Configurations

"DLD1 LTQ", "Mouse HCT" and "Yeast Velos" datasets configurations for ScanRanker evaluation:

peptide_mass_tolerance = 2.5
create_output_files = 1
ion_series = 0 1 1 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
fragment_ion_tolerance = 0.0
num_output_lines = 5
num_description_lines = 5
num_results = 500
show_fragment_ions = 0
print_duplicate_references = 1
enzyme_number = 0
diff_search_options = 15.9949 M
term_diff_search_options = 0.000 0.000
max_num_differential_AA_per_mod = 3
nucleotide_reading_frame = 0
mass_type_parent = 0
mass_type_fragment = 1
remove_precursor_peak = 0
ion_cutoff_percentage = 0.0
protein_mass_filter = 0 0
max_num_internal_cleavage_sites = 2
match_peak_count = 0
match_peak_allowed_error = 1
match_peak_tolerance = 1.0
add_C_Cysteine = 57.0215

X!Tandem Configurations

"DLD1 LTQ", "Mouse HCT" and "Yeast Velos" datasets configurations for ScanRanker evaluation:

protein, cleavage semi = yes
spectrum, search engine = tandem
spectrum, minimum cosine theta = 0.3
output, maximum valid expectation value = 1
residue, modification mass = 57.0215@C
residue, potential modification mass = 15.9949@M
protein, cleavage site = [RK]|{P}
spectrum, use contrast angle = no
list path, default parameters = iontrap.xml
output, xsl path = tandem-style.xsl
refine = no

output, results = all

PepNovo Configurations

model = CID_IT_TRYP
fragment_tolerance = 0.4 for LTQ-Velos and LTQ-Orbitrap, 0.15 for QSTAR
pm_tolerance = 2.5 for LTQ-Velos, 0.02 for LTQ-Orbitrap, 0.04 for QSTAR
no_quality_filter = true
num_solutions = 10
PTMs = C+57:M+16
use_spectrum_charge = false for LTQ-Velos, true for LTQ-Orbitrap and QSTAR
use_spectrum_mz = false for LTQ-Velos, true for LTQ-Orbitrap and QSTAR

TagRecon Configurations

“Histone Orbi” dataset configurations for ScanRanker evaluation:

PrecursorMzTolerance= 0.01
FragmentMzTolerance = 0.5
NTerminusMzTolerance = 0.5
CTerminusMzTolerance = 0.5
AdjustPrecursorMass = false
MaxPrecursorAdjustment = 1.008665
MinPrecursorAdjustment = -1.008665
PrecursorAdjustmentStep = 1.008665
NumSearchBestAdjustments = 3
DuplicateSpectra = true
UseChargeStateFromMS = true
NumChargeStates = 3
UseSmartPlusThreeModel = true
TicCutoffPercentage = 0.98
CleavageRules = "trypsin"
NumMaxMissedCleavages = 2
NumMinTerminiCleavages = 1
UseAvgMassOfSequences = false
StaticMods = ""
DynamicMods = "M ^ 15.9949 (Q * -17.026 ($ 42.015 C @ 57.021 [NQ] % 0.98"
MaxDynamicMods = 3
ExplainUnknownMassShiftsAs = "blindptms"
BlosumThreshold = -4
UseNETAdjustment = true
ComputeXCorr = true
MinCandidateLength = 5
MaxResults = 5

Pepitome Configurations

PrecursorMzToleranceRule = "avg"
MonoPrecursorMzTolerance = "10 ppm"
AvgPrecursorMzTolerance = "1.5 mz"
FragmentMzTolerance = "0.5 mz"
SpectrumListFilters = "peakPicking true 2-;chargeStatePredictor false 3 2 0.9"
RecalculateLibPepMasses = false
CleanLibSpectra = true
LibTicCuttoffPercentage = 0.98f
LibMaxPeakCount = 100
MonoisotopeAdjustmentSet = "0"
TicCutoffPercentage = 0.98
MaxPeakCount = 150
CleavageRules = "trypsin"
MaxMissedCleavages = 2
MinTerminiCleavages = 1
MinPeptideLength = 5
DynamicMods = "C % 57.021"
MaxDynamicMods = 3
StaticMods = ""
MaxResultRank = 2
FASTARefreshResults = false

ScanRanker Configurations

PrecursorMzTolerance = 1.25 for LTQ, 0.1 for LTQ-Orbitrap, 0.25 for QSTAR
FragmentMzTolerance = 0.5 for LTQ, 0.1 for LTQ-Orbitrap, 0.25 for QSTAR
IsotopeMzTolerance = 0.25 for LTQ and LTQ-Orbitrap, 0.125 for QSTAR
StaticMods = C 57.0215
NumChargeStates = 3
UseAvgMassOfSequences = 1 for LTQ, 0 for LTQ-Orbitrap and QSTAR
UseChargeStateFromMS = 0 for LTQ, 1 for LTQ-Orbitrap and QSTAR
UseMultipleProcessors = 0
WriteOutTags = 0

IDPicker Configurations

Maximum FDR = 0.05
Minimum distinct peptides = 2 (1 for synthetic peptide data)
Minimum additional peptides = 1
Minimum spectra per protein = 2 (1 for synthetic peptide data)

QuaMeter Configurations

RawDataPath = ../mzMLs/ # where to find the raw files for each idpDB

RawDataFormat = mzML # the file extension to expect for the raw files; e.g. mzML, mzXML, raw
Instrument = LTQ # if set to LTQ, average masses are used, else monoisotopic masses
ScoreCutoff = 0.05 # IDPicker FDR cutoff
ChromatogramMzLoIrOffset = 1.0mz # the loIr bound of the window for building chromatograms; can be in *m/z* or ppm
ChromatogramMzUpperOffset = 1.0mz # the upper bound of the window for building chromatograms; can be in *m/z* or ppm
ChromatogramOutput = false # if true, creates an mz5 file with the chromatograms (best vieId with SeeMS)

REFERENCES

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207

Arnett DR, Jennings JL, Tabb DL, Link AJ, Weil PA (2008) A Proteomics Analysis of Yeast Mot1p Protein-Protein Associations. Mol Cell Proteomics 7:2090 –2106

Baker PR, Medzihradszky KF, Chalkley RJ (2010) Improving Software Performance for Peptide Electron Transfer Dissociation Data Analysis by Implementation of Charge State- and Sequence-Dependent Scoring. Mol Cell Proteomics 9:1795–1803

Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24:1285–1292

Beer I, Barnea E, Ziv T, Admon A (2004) Improving large-scale proteomics by clustering of mass spectrometry data. Proteomics 4:950–960

Bereman MS, Canterbury JD, Egertson JD, Horner J, Remes PM, Schwartz J, Zabrouskov V, MacCoss MJ (2011) Evaluation of Front-End Higher Energy Collision-Induced Dissociation on a Benchtop Dual-Pressure Linear Ion Trap Mass Spectrometer for Shotgun Proteomics. Anal Chem 84:1533–1539

Bern M, Cai Y, Goldberg D (2007) Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. Anal Chem 79:1393–1400

Bern M, Goldberg D, McDonald WH, Yates III JR (2004) Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 20:i49

Bousquet-Dubouch M-P, Fabre B, Monsarrat B, Burlet-Schiltz O (2011) Proteomics to study the diversity and dynamics of proteasome complexes: from fundamentals to the clinic. Expert Rev Proteomics 8:459–481

Breukelen B van, Georgiou A, Drugan MM, Taouatas N, Mohammed S, Heck AJR (2010) LysNDeNovo: an algorithm enabling de novo sequencing of Lys-N generated peptides fragmented by electron transfer dissociation. Proteomics 10:1196–1201

Brosch M, Choudhary J (2010) Scoring and validation of tandem MS peptide identification methods. Methods Mol Biol 604:43–53

Brosch M, Yu L, Hubbard T, Choudhary J (2009) Accurate and Sensitive Peptide Identification with Mascot Percolator. J Proteome Res 8:3176–3181

Burgess EF, Ham A-JL, Tabb DL, Billheimer D, Roth BJ, Chang SS, Cookson MS, Hinton TJ, Cheek KL, Hill S, Pietenpol JA (2008) Prostate cancer serum

biomarker discovery through proteomic analysis of alpha-2 macroglobulin protein complexes. Proteomics Clin Appl 2:1223

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol Cell Proteomics 3:531–533

Chalkley R (2010) Instrumentation for LC-MS/MS in proteomics. Methods Mol Biol 658:47–60

Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, Fu Y, Yuan ZF, Wang HP, He SM, others (2010) pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra. Journal of Proteome Research 9:2713–2724

Choi H, Nesvizhskii AI (2008) Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. J Proteome Res 7:254–265

Chu F, Baker PR, Burlingame AL, Chalkley RJ (2010) Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. Molecular & Cellular Proteomics 9:25

Clauser KR, Baker P, Burlingame AL (1999) Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching. Anal Chem 71:2871–2882

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3:1454–1463

Coon JJ, Shabanowitz J, Hunt DF, Syka JEP (2005) Electron transfer dissociation of peptide anions. J Am Soc Mass Spectrom 16:880–882

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology 26:1367–1372

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. Journal of Proteome Research 10:1794–1805

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics:921

Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. J Proteome Res 5:1843–1849

Dasari S, Chambers MC, Martinez MA, Carpenter K, Ham A-J, Vega-Montoto L, Tabb DL (2012) Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. J Proteome Res

Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham A-JL, Tabb DL (2010) TagRecon: High-Throughput Mutation Identification through Sequence Tagging. J Proteome Res 9:1716–1726

Dayarathna MKDR, Hancock WS, Hincapie M (2008) A two step fractionation approach for plasma proteomics using immunodepletion of abundant proteins and multi-lectin affinity chromatography: Application to the analysis of obesity, diabetes, and hypertension diseases. J Sep Sci 31:1156–1166

Deutsch E (2008) mzML: A single, unifying data format for mass spectrometer output. Proteomics 8:2776–2777

Domon B (2006) Mass Spectrometry and Protein Analysis. Science 312:212–217

Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22:214–219

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214

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

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71

Fenyö D, Beavis RC (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 75:768–774

Finney GL, Blackler AR, Hoopmann MR, Canterbury JD, Wu CC, MacCoss MJ (2008) Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. Anal Chem 80:961–971

Flikka K, Martens L, Vandekerckhove J, Gevaert K, Eidhammer I (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. Proteomics 6:2086–2094

Frank AM (2009a) A Ranking-Based Scoring Function for Peptide−Spectrum Matches. J Proteome Res 8:2241–2252

Frank AM (2009b) Predicting Intensity Ranks of Peptide Fragment Ions. J Proteome Res 8:2226–2240

Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, Pevzner PA (2008) Clustering Millions of Tandem Mass Spectra. J Proteome Res 7:113–122

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

Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA (2007) De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res 6:114–123

Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem 78:5678–5684

Fukushima A, Kusano M, Redestig H, Arita M, Saito K (2009) Integrated omics approaches in plant systems biology. Current Opinion in Chemical Biology 13:532–538

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964

Gerster S, Qeli E, Ahrens CH, Bühlmann P (2010) Protein and Gene Model Inference Based on Statistical Modeling in K-Partite Graphs. PNAS 107:12101–12106

Görg A, Weiss W, Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4:3665–3685

Gstaiger M, Aebersold R (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet 10:617–627

Havilio M, Haddad Y, Smilansky Z (2003) Intensity-Based Statistical Scorer for Tandem Mass Spectrometry. Anal Chem 75:435–444

Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. PROTEOMICS 3:870–878

Hörth P, Miller CA, Preckel T, Wenz C (2006) Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics 5:1968–1974

Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443

Hubner NC, Ren S, Mann M (2008) Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. Proteomics 8:4862–4872

Johnson RS, Taylor JA (2002) Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. Molecular biotechnology 22:301–315

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4:923–925

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

Kenrick KG, Margolis J (1970) Isoelectric focusing and gradient gel electrophoresis: a two-dimensional technique. Anal Biochem 33:204–207

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

Kim S, Gupta N, Pevzner PA (2008) Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases. J Proteome Res 7:3354–3363

Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell 125:173–186

Klammer AA, MacCoss MJ (2006) Effects of modified digestion schemes on the identification of proteins from complex mixtures. J Proteome Res 5:695–700

Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS (2008) Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics 24:i348 –i356

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7:655–667

Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R (2010) Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. Molecular & Cellular Proteomics 9:1634

Licklider LJ, Thoreen CC, Peng J, Gygi SP (2002) Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. Anal Chem 74:3076–3083

Loecken EM, Dasari S, Hill S, Tabb DL, Guengerich FP (2009) The bis-electrophile diepoxybutane cross-links DNA to human histones but does not result in enhanced mutagenesis in recombinant systems. Chem Res Toxicol 22:1069–1076

Ma Z-Q, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL (2009) IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. J Proteome Res 8:3872–3881

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2337–2342

Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, Horning S (2006) Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. Anal Chem 78:2113–2120

Mann M, Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Analytical Chemistry 66:4390–4399

Matthiesen R, Trelle MB, H\ojrup P, Bunkenborg J, Jensen ON (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J Proteome Res 4:2338–2347

Moore RE, Young MK, Lee TD (2002) Qscore: an algorithm for evaluating SEQUEST database search results. Journal of the American Society for Mass Spectrometry 13:378–386

Na S, Jeong J, Park H, Lee K-J, Paek E (2008) Unrestrictive Identification of Multiple Post-Translational Modifications from Tandem Mass Spectrometry Using an Error-Tolerant Algorithm Based on an Extended Sequence Tag Approach. Mol Cell Proteomics 7:2452–2463

Na S, Paek E (2006) Quality Assessment of Tandem Mass Spectra Based on Cumulative Intensity Normalization. J Proteome Res 5:3241–3248

Nesvizhskii AI (2007) Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 367:87–119

Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteome Res 73:2092–2123

Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data. Molecular & Cellular Proteomics 4:1419

Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data. Molecular & Cellular Proteomics 5:652

Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Meth 4:787–797

Ning K, Fermin D, Nesvizhskii AI (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. Proteomics 10:2712–2718

Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M (2007) Higher-energy C-trap dissociation for peptide modification analysis. Nature Methods 4:709–712

Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8:2759–2769

Paizs B, Suhai S (2005) Fragmentation pathways of protonated peptides. Mass Spectrom Rev 24:508–548

Pan C, Park BH, McDonald WH, Carey PA, Banfield JF, VerBerkmoes NC, Hettich RL, Samatova NF (2010) A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. BMC bioinformatics 11:118

Paulovich AG, Billheimer D, Ham AJ., Vega-Montoto L, Rudnick PA, Tabb DL, Wang P, Blackman RK, Bunk DM, Cardasis HL, others (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. Mol Cell Proteomics 9:242

Perkins DN, Pappin DJ., Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Pernemalm M, Lewensohn R, Lehtiö J (2009) Affinity prefractionation for MS-based plasma proteomics. Proteomics 9:1420–1427

Qeli E, Ahrens CH (2010) PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol 28:647–650

Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham A-JL, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-

Montoto L, Variyath AM, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Carr SA, Fisher SJ, Gibson BW, Paulovich AG, Regnier FE, Rodriguez H, Spiegelman C, Tempst P, Liebler DC, Stein SE (2010) Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses. Mol Cell Proteomics 9:225 –241

Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, Zhang H, Zha X-M, Polakiewicz RD, Comb MJ (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol 23:94–101

Sadygov RG, Good DM, Swaney DL, Coon JJ (2009) A New Probabilistic Database Search Algorithm for ETD Spectra. J Proteome Res 8:3198–3205

Sadygov RG, Yates JR 3rd (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 75:3792–3798

Salmi J, Moulder R, Filén J-J, Nevalainen OS, Nyman TA, Lahesmaa R, Aittokallio T (2006) Quality classification of tandem mass spectrometry data. Bioinformatics 22:400–406

Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, Kuster B (2011) Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics 10:M110.003830

Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmström J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, Aebersold R, Mering C von, Hengartner MO (2009) Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. PLoS Biol 7:e1000048

Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, Lebleu VS, Duncan MB, Vander Heiden MG, Neveu JM, Lane WS, Cottrell JS, Horner JR, Cantley LC, Kalluri R, Asara JM (2009) Biomolecular Characterization and Protein Sequences of the Campanian Hadrosaur B. canadensis. Science 324:626–631

Searle BC (2010) Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. PROTEOMICS 10:1265–1269

Shen C, Wang Z, Shankar G, Zhang X, Li L (2008) A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. Bioinformatics 24:202

Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem 73:1917–1926

Slebos RJC, Brock JWC, Winters NF, Stuart SR, Martinez MA, Li M, Chambers MC, Zimmerman LJ, Ham AJ, Tabb DL, Liebler DC (2008) Evaluation of Strong Cation Exchange versus Isoelectric Focusing of Peptides for Multidimensional Liquid Chromatography-Tandem Mass Spectrometry. Journal of Proteome Research 7:5286–5294

Slotta DJ, McFarland MA, Markey SP (2010) MassSieve: Panning MS/MS peptide data for proteins. PROTEOMICS 10:3035–3039

Spivak M, Tomazela D, Weston J, MacCoss MJ, Noble WS (2011) Direct Maximization of Protein Identifications from Tandem Mass Spectra. Mol Cell Proteomics

States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol 24:333–338

Sun RX, Dong MQ, Song CQ, Chi H, Yang B, Xiu LY, Tao L, Jing ZY, Liu C, Wang LH, others (2010) Improved Peptide Identification for Proteomic Analysis Based on Comprehensive Characterization of Electron Transfer Dissociation Spectra. Journal of Proteome Research

Swaney DL, McAlister GC, Coon JJ (2008) Decision tree–driven tandem mass spectrometry for shotgun proteomics. Nat Meth 5:959–964

Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. Journal of proteome research 9:1323–1329

Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6:654–661

Tabb DL, Friedman DB, Ham A-JL (2006) Verification of automated peptide identifications from proteomic tandem mass spectra. Nat Protoc 1:2213–2222

Tabb DL, Ma Z-Q, Martin DB, Ham A-JL, Chambers MC (2008) DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res 7:3838–3846

Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. Anal Chem 75:2470–2477

Tabb DL, Thompson MR, Khalsa-Moyers G, VerBerkmoes NC, McDonald WH (2005) MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. J Am Soc Mass Spectrom 16:1250–1261

Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T, Matsuo T (1988) Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2:151–153

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

Thingholm TE, Jensen ON, Larsen MR (2009) Enrichment and separation of mono- and multiply phosphorylated peptides using sequential elution from IMAC prior to mass spectrometric analysis. Methods Mol Biol 527:67–78, xi

Trnka MJ, Burlingame AL (2010) Topographic Studies of the GroEL-GroES Chaperonin Complex by Chemical Cross-linking Using Diformyl Ethynylbenzene. Molecular & Cellular Proteomics 9:2306

Wang L, Li D-Q, Fu Y, Wang H-P, Zhang J-F, Yuan Z-F, Sun R-X, Zeng R, He S-M, Gao W (2007) pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. Rapid Commun Mass Spectrom 21:2985–2991

Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology 19:242–247

Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3:179–196

Wilhelm M, Kirchner M, Steen JAJ, Steen H (2012) mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets. Mol Cell Proteomics 11

Witze ES, Old WM, Resing KA, Ahn NG (2007) Mapping protein post-translational modifications with mass spectrometry. Nat Methods 4:798–806

Wysocki VH, Resing KA, Zhang Q, Cheng G (2005) Mass spectrometry of peptides and proteins. Methods 35:211–222

Wysocki VH, Tsaprailis G, Smith LL, Breci LA (2000) Mobile and localized protons: a framework for understanding peptide dissociation. Journal of Mass Spectrometry 35:1399–1406

Xu H, Freitas MA (2008) Monte Carlo Simulation-Based Algorithms for Analysis of Shotgun Proteomic Data. J Proteome Res 7:2605–2615

Xu M, Geer LY, Bryant SH, Roth JS, Kowalak JA, Maynard DM, Markey SP (2005) Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. J Proteome Res 4:300–305

Yadav AK, Kumar D, Dash D (2011) MassWiz: A Novel Scoring Algorithm with Target-Decoy Based Analysis Pipeline for Tandem Mass Spectrometry. J Proteome Res 10:2154–2160

Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, Chen X, Markey SP, Kowalak JA (2004) DBParser: web-based software for shotgun proteomic data analyses. J Proteome Res 3:1002–1008

Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 11:49–79

Zhang Z (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 76:3908–3922

Zhang Z (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem 77:6364–6373

Zhang B, Chambers MC, Tabb DL (2007) Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency. J Proteome Res 6:3549–3557

Zhang W, Li F, Nie L (2010) Integrating Multiple “omics” Analysis for Microbial Biology: Application and Methodologies. Microbiology 156:287–301

Zubarev R, Mann M (2007) On the proper use of mass accuracy in proteomics. Mol Cell Proteomics 6:377–381