INTELLIGENT DATA ACQUISITION FOR MASS SPECTROMETRY-BASED PROTEOMICS

by

Derek J. Bailey

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

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 1/23/14

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Chemistry
Lloyd M. Smith, Professor, Chemistry
Michael R. Sussman, Professor, Biochemistry
David J. Pagliarini, Assistant Professor, Biochemistry
Sushmita Roy, Assistant Professor, Biostatistics

To Becky, my wife.

ACKNOWLEDGMENTS

> *Ideally, the end of extrinsically applied education should be the start of an education that is motivated intrinsically. At that point the goal of studying is no longer to make the grade, earn a diploma, and find a good job. Rather, it is to understand what is happening around one, to develop a personally meaningful sense of what one's experience is all about. From that will come the profound joy of the thinker.*
>
> — Mihaly Csikszentmihalyi (1990)

To those who continually encouraged me to love learning and thinking, without your support this work, and who I am today, would not exist.

I start with sincere thanks and appreciation for Joshua Coon, my research advisor. His guidance and mentorship these past years have truly made me a better scientist, writer, and communicator. With access to world-class science, the newest instruments, and a plethora of professional contacts, working in his research group has been extremely helpful in my development. The opportunities he provided me have truly propelled my career forward, and I am confident the training I received from him will continue to help me down the road.

In addition to Joshua's direct support, he has also assembled a great team of scientists and fellow graduate students that I have worked with. First and foremost would be our staff scientist Mike Westphall, who has provided incredible amounts of wisdom to me. If anyone deserved the title of advisor, it would be him. Alex Hebert and Christopher Rose receive special thanks for listening to all my crazy ideas and all their scientific advice they provided. To the other graduate students: Catie Minogue, Anna Merrill, Greg Potts, Alicia Richards, Allison Balloon, Nick Riley, Nick Kwiecien, Arne Ulbrich, Tim Rhoads and Molly McDevitt, thank you for your support. Additionally, these former graduate students had a special part in my early years in the lab: Graeme McAlister, Doug Phanstiel, Violet Lee, Justin Brumbaugh, Aaron Ledvina, Amelia Peterson, Jason Russell, and Danielle Swaney. I especially want to thank Craig Wenger, whose expertise in computer related science is something I try to achieve.

Before my graduate work at Madison, there was my undergraduate education at Wheaton College, IL. And no two people, other than my parents, have had a bigger impact on my life than Daniel and Lisa Burden. It is under their mentorship and friendship that I first started a career in science and developed the skills to realize my dreams. Their willingness to individually train and guide students at Wheaton has surely touched more lives than just mine, but I am indebted to their commitment to me and my future. I strive to emulate their love of science, God, and for each other; they are truly people to model.

Without the strong and constant support of my family, all of this would be impossible. My parents have constantly provided for me, their support and love are without equal. Their encouragement to love learning, reading, the outdoors, and games have made me who I am today.

Lastly, to my wife Rebecca, who has seen my graduate school journey first hand. I love you and the companionship we share, advice you provide, and the smile you bring me. Thank you for the life we had together and the future to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AC	Alternating current
AGC	Automatic gain control
API	Application programming interface
a.u.	Arbitrary unit
BCA	Bicinchoninic acid protein assay
BLAST	Basic local alignment search tool
CAD	Collision-activated dissociation
CEO	Calculated elution order
CI	Chemical ionization
CID	Collision-induced dissociation
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
C#	C sharp, a programming language
CSMSL	C# Mass Spectrometry Library
Da	Dalton, the atomic mass unit
DC	Direct current
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid

DT	Decision tree
DTT	Dithiothreitol
ECD	Electron-capture dissociation
EO	Elution order
EOA	Elution order algorithm
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
E-value	Expectation value
FAB	Fast-atom bombardment
FASTA	A format for storing protein sequences
FDR	False discovery rate
FT	Fourier transform
FT-ICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
GUI	Graphical user interface
HCD	Higher-energy collisional dissociation
HPLC	High-performance liquid chromatography
HTCondor	High throughput condor
Hz	Hertz, inverse seconds

IDA	Intelligent data acquisition
INC	Inclusion list
IP	Intellectual property
inSeq	Instant sequencing algorithm
IT	Ion trap
ITCL	Ion trap control language
LC	Liquid chromatography
m	Mass
MALDI	Matrix-assisted laser desorption-ionization
min	Minute
MS	Mass spectrometry
MS^1	Survey mass analysis
MS^n	Tandem mass spectrometry
MS/MS	Tandem mass spectrometry
NCE	Normalized collision energy
m/z	Mass-to-charge ratio
nLC	Nanoflow liquid chromatography
OMSSA	Open Mass Spectrometry Search Algorithm
ppm	Part per million

PRM	Parallel reaction monitoring
PSM	Peptide-spectrum match
PTM	Post-translational modification
QM	QuantMode
RP	Reverse phase
RT	Retention time
s	Second
SDF	Site-determining fragment
SILAC	Stable isotope labeling by amino acids in cell culture
SIM	Single ion monitoring
S/N	Signal-to-noise ratio
SRM	Selected reaction monitoring
SCX	Strong-cation exchange
S/N	Signal-to-noise ratio
Th	Thomson, the unit of the mass-to-charge ratio
TIC	Total-ion chromatogram
TMT	Tandem mass tag
XIC	Extracted-ion chromatogram
z	Charge

ABSTRACT

The following chapters detail improvements made to the data acquisition methods used in mass spectrometry-based proteomics. Mass spectrometers were empowered to analyze spectral data in real-time and make informed decisions on how to proceed based on the results. Chapter 1 begins with a broad overview of proteomics, mass spectrometry, and commonly used data acquisition methods. The second chapter focuses on the instant sequencing (*inSeq*) algorithm that improves reproducibility, boosts quantitative results, and localizes more post-translational modifications (PTMs) in LC-MS/MS experiments. Chapter 3 introduces a novel real-time elution ordering algorithm (EOA) to target hundreds of peptides across multiple LC-MS/MS experiments reproducibly. In chapter 4, two large software packages (COMPASS & CSMSL), which handle large-scale data analysis, are presented. The final chapter suggests future improvements to further IDA methods.

Chapter 1

PROTEOMIC ACQUISITION STRATEGIES FOR MASS SPECTROMETRY

Proteomics and Mass Spectrometry

Proteomics. Proteomics is the large-scale study of proteins, which are an essential part of life. If genes are the blueprints for life, then proteins are the construction workers, building supplies, and tools that empower and sustain life. Their involvement in life ranges from the diseases and aliments that cause impairment, to the therapeutics and medicines that cure them. From agriculture and food that provides energy, to the composition and structure of cells, there is very little in life that proteins do not affect. Understanding their role and function in biological systems is an overarching goal of life sciences.

Proteins are structurally similar to deoxyribonucleic acid (DNA); both are long, linear chemical polymers comprised of different monomers in a sequence. DNA has four nucleotide bases (G,T,A,C) and proteins are made up of roughly twenty amino acids (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y). The sequential order of these monomers in both DNA and proteins encodes information. The information stored in DNA is transcribed to mRNA and translated into proteins; DNA can be thought of as an instruction book, or blueprint for life. It provides the recipes to make all the proteins in the cell. On the other hand, the sequence of amino acids in proteins isn't used to store instructions, but rather it defines their structure. Proteins fold into complex three dimensional structures depending on their amino acid sequence and it is these 3D structures that provide the different mechanical and chemical functions for life to work.

A single organism can contain hundreds or thousands of different protein sequences depending on its complexity, and that set of proteins is called the proteome. The proteome of baker's yeast (*Saccharomyces cerevisiae*) contains ~4,600 different protein sequences. The human (*Homo sapiens*) proteome is much larger, with nearly 12,000 unique protein sequences that are known to be expressed. The interactions between proteins and other molecules within the cell provides many of the critical functions for life to exist and procreate. Proteins are essential to most cellular processes, and when they fail to adequately perform their duty, they can cause serious problems, even death. Thus it is very important to understand organisms' proteomes and how they are affected by different treatments and conditions. Understanding how proteins change in abundance and are covalently modified by chemical signals (e.g., phosphorylation, acetylation, ubiquitination, etc.), and alter their interactions with each other are important parts of the scientific field of study called 'systems biology'. Studying large and complex systems, such as the cell, relies on the identification and quantitation of large portions of a organism's proteome. However, detecting thousands of proteins and measuring their abundances is a daunting task. Only in the last two decades have technologies been developed that are capable of identifying and analyzing such large sets of proteins, and more often than not, the technology of choice is mass spectrometry (MS).

Mass Spectrometry. Mass spectrometry is a powerful analytical tool for measuring the mass of molecules. Current mass spectrometers can regularly measure mass to a single dalton (Da) or lower and are sensitive enough to detect as little as a few thousands molecules. Since each amino acid has a different mass, mass spectrometry is an ideal analytical technique to study and identify proteins. A mass spectrometer is an instrument that measures the mass of ions and is comprised of three parts: 1) an ionization source, 2) a mass analyzer, and 3) a detector.

Gas phase ions (negative or positive) are first generated by an ionization source from an analyte in either the solid or liquid phases. There are many ionization techniques in use today: hard-ionization methods such as electron impact ionization (EI) causes the analyte to fragment during ionization. Softer methods that minimize fragmentation include fast-atom bombardment (FAB)[1], matrix assisted laser desorption ionization (MALDI)[2,3], electrospray ionization (ESI)[4], and chemical ionization (CI)[5], among others. ESI has become the primary method for large-scale proteomic studies because of the ease which it can be coupled to a front-end separation technique such as liquid chromatography (LC). An ESI emitter is constructed at the end of a LC column, and ionizes the eluting proteins into the MS (LC-MS). Separation is often needed in large-scale proteomic experiments because samples can consist of overly complex mixtures of thousands of analytes, and separating them prior to ionization increases sensitivity.

The second part of a mass spectrometer is the mass analyzer, which separates the gas phase ions based on their mass-to-charge ratios (*m/z*). To effectively manipulate ions in the gas phase, various electrical devices have be developed to store, move, and analyze them. Since ions are charged particles, only their *m/z* ratio is detected, and their mass (*m*) is not directly measured. However, the mass can be calculated from the *m/z* if the number of charges is known. There are many different types of mass analyzers: magnetic sector, time-of-flight (TOF)[6], quadrupole mass filters[7], ion-traps[8], Orbitraps[9], and fourier transform ion cyclotron resonsnace (FT-ICR)[10]. Each analyzer relies on a different principle to separate ions by *m/z*, and some are able to separate better and/or faster than others. The Orbitrap and FT-ICR are capable of separating closely spaced *m/z* ions (i.e., high resolution), while ion-traps and TOFs can mass analyze very quickly with sufficient resolution. Each analyzer has its pros and cons, so it has become common to include multiple mass analyzers in one instrument to increase the capabilities of the MS. These multi-analyzer instruments are called hybrid instruments as they blend multiple techniques into one package.

The final portion of a mass spectrometer is the detector. Following mass separation, the ions need to be detected and converted into electrical signals to be recorded. Some detectors are destructive, consuming the ion when detected. Examples of these include Faraday cups and electron multipliers. Some detectors are non-destructive, being able to detect the ions without consuming them. These detectors are unique in that they can act both as a mass analyzer and detector. The Orbitrap and FT-ICR are examples of non-destructive, inductive detectors where an AC current is produced by ions oscillating within the detector. This generates an AC current in the detector that is stored and subsequently Fourier transformed into a *m/z* spectrum. These inductive detectors rely on Fourier transformation which scales with the length of acquisition. Thus, the longer the ions are detected the higher the resolution and increased signal-to-noise (S/N) achieved.

Mass spectrometers are powerful instruments that ionize, separate, and detect various analytes. While determining the mass of analytes is useful in and of itself, mass analysis of proteins is more convoluted. This is because the order of amino acids in a protein does not change the total mass. Two functionally different proteins could have the same amino acid contents, but in a different order, and would appear at the same *m/z* value. To ascertain their sequence—and determine their identity, additional analysis steps are needed. First, a set of ions are injected into the MS and a full scan is taken (MS or MS1). Then a certain *m/z* feature is isolated from the other ions in the MS. These isolated ions are then dissociated into smaller pieces (fragments) and mass analyzed again (MS/MS). Proteins can be dissociated by a variety of different methods. The most common approach is to forcefully collide the ions with background gas molecules that are present in the instrument (Collision-Activated Dissociation, CAD). Other approaches exist, such as electron-capture dissociation (ECD)[11] and electron-transfer dissociation (ETD)[12], as well as infrared mulitple photon dissociation (IRMPD)[13] and higher energy C-trap dissociation (HCD)[14]. Following dissociation, a fragmentation spectrum is collected and can help provide clues on the sequence of the protein being analyzed. This process is called tandem mass spectrometry, as multiple mass anlaysis steps are taken to identify proteins. While tandem mass analysis of intact proteins is possible, their large size, complex fragmentation spectra, and poor separability make large-scale analysis of proteins challenging. One popular approach called bottom-up proteomics alleviates some of these issues by breaking proteins apart into smaller pieces and mass analyzing those.

Acquisition Methods for Shotgun Proteomics

Shotgun proteomics is the process of digesting proteins into smaller pieces, called peptides, prior to separation and mass analysis (LC-MS). This scheme (bottom-up proteomics) offers many benefits over mass analysis of intact proteins (top-down proteomics). First, the smaller size of peptides makes separation by LC simpler. Since peptides have less 3D structure than proteins, they interact more with the stationary phase of the separation, which helps improves the separation. Differences in ionization between peptides and proteins is another major factor. Peptides often ionize into a smaller number of charge states than proteins because they contain fewer charge-carrying sites. This concentrates the signal into fewer states, increasing the S/N for any given one. Proteins often exist in dozens of different charge states, which dilutes the S/N among them. On top of the charge state distribution, the wide isotopic distributions of proteins further decreases the S/N and increases spectral complexity. The distribution of *m/z* for peptides is also centered in the optimal mass range of most mass spectrometers (e.g., 300 - 1500 *m/z*). Finally, the smaller the analyte the less complex the fragmentation spectra are and that usually means easier interpretation and identification. These and other reasons make peptide mass analysis easier than intact protein mass analysis.

While bottom-up proteomics offers many advantages compared to top-down analysis, new challenges also arise. First, the sample becomes much more complex. For example, the proteome of yeast contains ~4,600 proteins, but when they are digested into peptides by proteases, such as trypsin, hundreds of thousands of peptides result. Another issue that surfaces is increased ambiguity in protein identification. Following sequencing of peptides by LC-MS/MS, computer algorithms map these peptides back to their parent proteins. But shorter peptide sequences are more likely to have originated from more than one protein, obscuring which protein was actually identified. This can be partially combated by identifying other peptides from those proteins to help distinguish them apart. Other challenges are also present, but most proteomic publications make use of the shotgun proteomic scheme.

The typical shotgun proteomic workflow begins with protein extraction from cell cultures or tissues. Proteins are then isolated and digested into peptides by proteolysis and are separated by liquid chromatography. Peptides elute from the LC column and are ionized by ESI into the MS. From here a variety of acquisition methods are can be used to generate tandem mass spectra (MS/MS) of the eluting peptides; these will be discussed in detail below. Following acquisition, the collection of spectra are searched against a database of protein sequences to identify the peptide sequences from the fragmentation spectra. This is followed by statistical analysis that culls out false identifications.[15] Finally, the identified peptides are assembled back into protein groups and optional quantitative analysis can be conducted. This results in a list of identified proteins, their relative abundances, and possible post-translational modifications (PTMs). This workflow first emerged in 2001 and has changed very little since.[16] Improvements to each part of this worklow, especially the advent of new hybrid mass spectrometers, have propelled this methodology from identifying 1,483 yeast proteins in ~68 hours to >4,000 proteins in an hour, all in the past decade.[17]

Probably the one aspect of this workflow that has changed the least is the data acquisition method—how the MS decides what *m/z* to dissociate and MS/MS analyze. The two overarching methods for data acquisition in a shotgun experiment are the data-dependent and data-independent methods. Data-dependent acquisition (DDA) relies on surveying all the intact *m/z* of peptide precursors with a full MS scan, followed by successive isolation and fragmentation of different *m/z* peaks based on their intensity.[18,19] Data-independent acquisition (DIA), on the other hand, forgoes the initial survey scan and iteratively isolates and fragments different *m/z* regions in a predictable fashion.[20,21]

Data-Dependent Acquisition Data-dependent acquisition relies on gathering information about the peptides currently eluting to select the best candidates for MS/MS sampling. It is typically performed with an initial MS scan to mass analyze all the peptide precursors. Then the mass spectrometer selects the top N most intense *m/z* features to undergo dissociation and MS/MS analysis in subsequent scans. The value of N is typically between 5 and 20 depending on the instrument. Following acquisition of N MS/MS spectra, the whole process is repeated with another survey scan. This straightforward and simple method is highly effective and has been relatively unchanged since its debut.

The DDA method has been supplemented with additional filtering criteria to improve the diversity of *m/z* peaks sampled. With the dynamic exclusion filter, closely-spaced *m/z* features are excluded from being reselected within some time range since it was first selected (e.g., 30 seconds). This prevents the repeated sampling and identification of the same precursor and tries to sample a wider population of precursors. Other DDA filtering criteria try to avoid sampling precursors that will not produce identifiable fragmentation spectra. With high-resolution spectra, *m/z* features that are singly charged are often avoided as they produce poor MS/MS spectra. Other filters look at the isotopic ratios of the analytes and avoid sampling features that don't display peptidic ratios. These and other filters help DDA segment the MS time to select a diverse set of peptide precursors during an experiment with high throughput.

One of the biggest issues with DDA is that the analyte must be detected in the initial survey scan in order to be sampled and identified. If a precursor, for whatever reason, never exceeds a S/N threshold to be selected, it will never be identified. This makes DDA a stochastic process, depending on the quality of the survey scan data to make its future decisions. This leads to irreproducible sampling and identification across multiple experiments, leaving datasets incomplete. The other acquisition method, Data-independent acquisition, tries to avoid this issue by skipping the survey scan altogether.

Data-Independent Acquisition Data-independent acquisition is a methodology where the MS iteratively isolates and dissociates different *m/z* regions regardless of detection in a survey scan. There are a few different approaches that DIA methods can take. The simplest is to repetitively isolate and dissociate the same *m/z* region for the entire LC separation. This guarantees that any peptide precursor whose *m/z* is within the isolation range will be analyzed. This method is extremely low-throughput though, only able to analyze a few dozen different precursors over the course of the separation.

Increased throughput can be gained by changing the isolated *m/z* region throughout the separation. For example, in the first MS/MS scan the *m/z* range 300 - 303 is isolated and mass analyzed. The following scan the *m/z* range changes to 303 - 306, the third scan to 306 - 309, and so forth. After the complete mass range is analyzed once (e.g., 300-1200 *m/z*), it starts over at 300 - 303 *m/z*. This approach samples all possible *m/z* ranges during the experiment, but each individual *m/z* range must wait until all the other ones have been sampled. This leads to long times between analysis and could lead to completely missing an eluting analyte from being sampled. However, when not limited by time or sample amounts this approach can produce a comparable number of identifications to DDA methods.[22]

Probably the most popular DIA method is selected reaction monitoring (SRM).[23,24] This type of experiment is usually conducted on a triple-quadrupole instrument in a scheduled fashion. Scheduling involves breaking up the LC-MS/MS experiment into different time segments, and targeting a subset of peptides per segment. In an SRM experiment, the first quadrupole isolates a single precursor during its expected elution time, the second quadrupole fragments those ions, and the final quadrupole isolates another *m/z* before the ions reach the detector. SRM is a very sensitive and selective method, and is the gold-standard for targeted proteomics. The method, however, is low-throughput and is able to target only a few hundred peptides per LC-MS experiment. SRM methods also do not usually include a full MS/MS spectrum for each peptide, favoring increased sensitivity at the cost of more ambiguity. Recent work by our lab and others have extended the SRM method to work on high-resolution instruments with a full MS/MS scan being acquired, a method called parallel reaction monitoring (PRM).[25,26]

More recent advances in DIA methods include isolating and mass analyzing large *m/z* regions. Here, instead of 3 *m/z* isolation ranges, the isolation window is open up to 10, 20 or even 50 *m/z*. Methods such as ”sequential windowed data independent acquisition of the total high-resolution mass spectra” (SWATH) and MS everything (MS^E) seek to further improve the throughput and reproducibility of DIA methods with informatic advances.[27,28] Here multiple precursors are co-isolated and co-fragmented, producing complex fragmentation spectra that are deconvoluted post-acquisition. This increases throughput as less *m/z* regions need to be analyze per cycle. However, post-translational modification (PTM) analysis is stymied because the complex fragmentation spectra which are produced make PTM localization a very difficult process.

Intelligent Data Acquisition

Although most mass spectrometry-related technologies have constantly improved over the past two decades, improvements to the data acquisition methods have progressed at a far slower pace. The DDA strategy has only been modestly supplemented in the past decade while the DIA strategy seeks to remove all intelligence from the acquisition in favor of simplicity. Perhaps the biggest step forward in increasing the performance of DDA methods was the introduction of the ETD-CAD decision tree (DT) algorithm.[29] Here, precursors were either dissociated with ETD or CAD based on their *m/z* and charge-state, increasing identification rates. Unlike DDA, the DT algorithm incorporated multiple pieces of data (*m/z*, *z*, intensity, etc.) to make an advanced decision.

Following the advent of the DT algorithm, which relied on advanced analysis of a MS scan, we thought it could be expanded to the analysis of MS/MS scans. Direct analysis of MS/MS scans in real-time provides a lot more information that could be used to make more advanced decisions. The following chapters describe some of the first work on improving the intelligence of MS acquisition methods for proteomic research. These methods, grouped under the term intelligent data acquisition (IDA), represent new methodologies in how the MS selects peptide precursors and improves the data quality. In chapter 2, the development of the first real-time database searching algorithm (*inSeq*) is described. The MS was empowered to identify MS/MS spectra immediately following acquisition, allowing for the MS to decide what to do next. Improvements to reproducibility, quantitative accuracy, and PTM analysis are demonstrated over traditional DDA methods. Chapter 3 summarizes our real-time algorithms for improving the run-to-run reproducibility of peptide identification. Peptides are scheduled based on their relative elution order instead of the more typical absolute retention times. The IDA method could determine the overall elution order by analyzing survey MS scans, and then subsequently target peptides in a DIA fashion. The method increased the number of peptides identified in repeated experiments by over 50%. Chapter 4 is devoted to the programming and software frameworks used for data analysis and real-time control. The chapter starts off with a summary of the Coon OMSSA Proteomic Analysis Software Suite (COMPASS) and the improvements made to it since its initial publication.[30] The suite is a complete data-analysis package for tandem mass spectrometry data. The chapter continues with a brief discussion on the developments of the C# Mass Spectrometry Library (CSMSL). This library provides many tools and programs to develop new analysis programs in a fast and easy manner. The final chapter in this document looks at the future of IDA methods and what challenges they face and offers a few suggestions on possible solutions.

References

[1] M. Barber, R. S. Bordoli, R. D. Sedgwick, and A. N. Tyler, “Fast atom bombardment of solids (fab) - a new ion-source for mass-spectrometry,” *Journal of the Chemical Society-Chemical Communications*, no. 7, pp. 325–327, 1981.

[2] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, and T. Yoshida, “Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry,” *Rapid Commun. Mass Spectrom.*, vol. 2, no. 8, pp. 151–153, 1988.

[3] M. Karas and F. Hillenkamp, “Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons,” *Analytical Chemistry*, vol. 60, no. 20, pp. 2299–2301, 1988.

[4] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, “Electrospray ionization for mass-spectrometry of large biomolecules,” *Science*, vol. 246, no. 4926, pp. 64–71, 1989.

[5] M. S. B. Munson and F. H. Field, “Chemical ionization mass spectrometry .i. general introduction,” *Journal of the American Chemical Society*, vol. 88, no. 12, p. 2621, 1966.

[6] W. Stephens, “A pulsed mass spectrometer with time dispersion,” in *Am. Phys. Soc.*, vol. 21, p. 22.

[7] W. Paul and H. Steinwedel, “Ein neues massenspektrometer ohne magnetfeld,” *Z. Naturforschg.*, vol. 8, no. a, p. 448, 1953.

[8] G. C. Stafford, P. E. Kelley, J. E. P. Syka, W. E. Reynolds, and J. F. J. Todd, “Recent improvements in and analytical applications of advanced ion trap technology,” *International Journal of Mass Spectrometry and Ion Processes*, vol. 60, no. Sep, pp. 85–98, 1984.

[9] A. Makarov, “Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis,” *Anal Chem*, vol. 72, no. 6, pp. 1156–62, 2000.

[10] Comisaro.Mb and A. G. Marshall, “Fourier-transform ion-cyclotron resonance spectroscopy,” *Chemical Physics Letters*, vol. 25, no. 2, pp. 282–283, 1974.

[11] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty, “Electron capture dissociation of multiply charged protein cations. a nonergodic process,” *Journal of the American Chemical Society*, vol. 120, no. 13, pp. 3265–3266, 1998.

[12] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, “Peptide and protein sequence analysis by electron transfer dissociation mass spec-

trometry," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9528–9533, 2004.

[13] D. P. Little, J. P. Speir, M. W. Senko, P. B. Oconnor, and F. W. Mclafferty, "Infrared multiphoton dissociation of large multiply-charged ions for biomolecule sequencing," *Analytical Chemistry*, vol. 66, no. 18, pp. 2809–2815, 1994.

[14] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, "Higher-energy c-trap dissociation for peptide modification analysis," *Nature Methods*, vol. 4, no. 9, pp. 709–712, 2007.

[15] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nat Methods*, vol. 4, no. 3, pp. 207–14, 2007.

[16] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology," *Nat Biotechnol*, vol. 19, no. 3, pp. 242–7, 2001.

[17] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Mol Cell Proteomics*, vol. 13, no. 1, pp. 339–47, 2014.

[18] A. R. Dongre, J. K. Eng, and J. R. Yates, "Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins," *Trends Biotechnol*, vol. 15, no. 10, pp. 418–25, 1997.

[19] A. Ducret, I. Van Oostveen, J. K. Eng, J. R. Yates, and R. Aebersold, "High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry," *Protein Sci*, vol. 7, no. 3, pp. 706–19, 1998.

[20] S. Purvine, J. T. Eppel, E. C. Yi, and D. R. Goodlett, "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer," *Proteomics*, vol. 3, no. 6, pp. 847–50, 2003.

[21] J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nat Methods*, vol. 1, no. 1, pp. 39–45, 2004.

[22] M. R. Hoopmann, G. E. Merrihew, P. D. von Haller, and M. J. MacCoss, "Post analysis data acquisition for the iterative ms/ms sampling of proteomics mixtures," *J Proteome Res*, vol. 8, no. 4, pp. 1870–5, 2009.

[23] A. Wolf-Yadlin, S. Hautaniemi, D. A. Lauffenburger, and F. M. White, "Multiple reaction monitoring for robust quantitative proteomic analysis of cellular

signaling networks," *Proc Natl Acad Sci U S A*, vol. 104, no. 14, pp. 5860–5, 2007.

[24] V. Lange, P. Picotti, B. Domon, and R. Aebersold, "Selected reaction monitoring for quantitative proteomics: a tutorial," *Mol Syst Biol*, vol. 4, p. 222, 2008.

[25] S. E. Ong and M. Mann, "Mass spectrometry-based proteomics turns quantitative," *Nat Chem Biol*, vol. 1, no. 5, pp. 252–62, 2005.

[26] S. Gallien, E. Duriez, C. Crone, M. Kellmann, T. Moehring, and B. Domon, "Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer," *Mol Cell Proteomics*, vol. 11, no. 12, pp. 1709–23, 2012.

[27] L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis," *Mol Cell Proteomics*, vol. 11, no. 6, p. O111 016717, 2012.

[28] R. S. Plumb, K. A. Johnson, P. Rainville, B. W. Smith, I. D. Wilson, J. M. Castro-Perez, and J. K. Nicholson, "Uplc/ms(e); a new approach for generating molecular fragment information for biomarker structure elucidation," *Rapid Commun Mass Spectrom*, vol. 20, no. 13, pp. 1989–94, 2006.

[29] D. L. Swaney, G. C. McAlister, and J. J. Coon, “Decision tree-driven tandem mass spectrometry for shotgun proteomics,” *Nature Methods*, vol. 5, no. 11, pp. 959–964, 2008.

[30] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, “Compass: a suite of pre- and post-search proteomics software tools for omssa,” *Proteomics*, vol. 11, no. 6, pp. 1064–74, 2011.

Chapter 2

INSTANT SPECTRAL ASSIGNMENT FOR ADVANCED DECISION TREE-DRIVEN MASS SPECTROMETRY

Summary

We have developed and implemented a sequence identification algorithm (*inSeq*) that processes tandem mass spectra in real-time using the mass spectrometer's (MS) on board processors. *inSeq* relies on accurate mass tandem MS data for swift spectral matching with high accuracy. The instant spectral processing technology takes ~16 ms to execute and provides information to enable autonomous, real-time decision making by the MS system. Using *inSeq*, and its advanced decision tree (DT) logic, we demonstrate: (1) real-time prediction of peptide elution windows en masse (~3 minute width, 3,000 targets), (2) significant improvement of quantitative precision and accuracy (~3X boost in detected protein differences), and (3) boosted rates of post-translational modification (PTM) site localization (90% agreement in real-time vs. offline localization rate and a ~25% gain in localized sites). The DT logic enabled by *inSeq* promises to circumvent longstanding problems with the conventional data-dependent acquisition paradigm and provides a direct route to

streamlined and expedient targeted protein analysis.

Introduction

The shotgun sequencing method has rapidly evolved over the past two decades.[1,2] In this strategy eluting peptide cations have their mass-to-charge (*m/z*) values measured in the MS scan. Then precursor *m/z* values are selected for a series of sequential tandem MS events (MS/MS). This succession is cycled for the duration of the analysis. The process, called data-dependent acquisition (DDA), is at the very core of shotgun analysis and has not changed for over fifteen years; MS hardware, however, has. Major improvements in MS sensitivity, scan rate, mass accuracy and resolution have been achieved. Orbitrap hybrid systems, for example, routinely achieve low ppm mass accuracy with MS/MS repetition rates of 5-10 Hz.[3,4] Constant operation of such systems generates hundreds of thousands of spectra in days. These MS/MS spectra are then mapped to sequences using database search algorithms.[5–7]

The DDA sampling strategy offers an elegant simplicity and has proven highly useful for discovery-driven proteomics. Of recent years, however, emphasis has shifted from identification to quantification—often with certain targets in mind. In this context faults in the DDA approach have become increasingly evident.

There are two primary limitations of the DDA approach: First, is poor run-to-run reproducibility and, second, is the inability to effectively target peptides of interest.[8] Hundreds of peptides often co-elute so that low-level signals often get selected in one run and not the next. And selecting *m/z* peaks to sequence by abundance certainly does not offer the opportunity to inform the system of pre-selected targets.

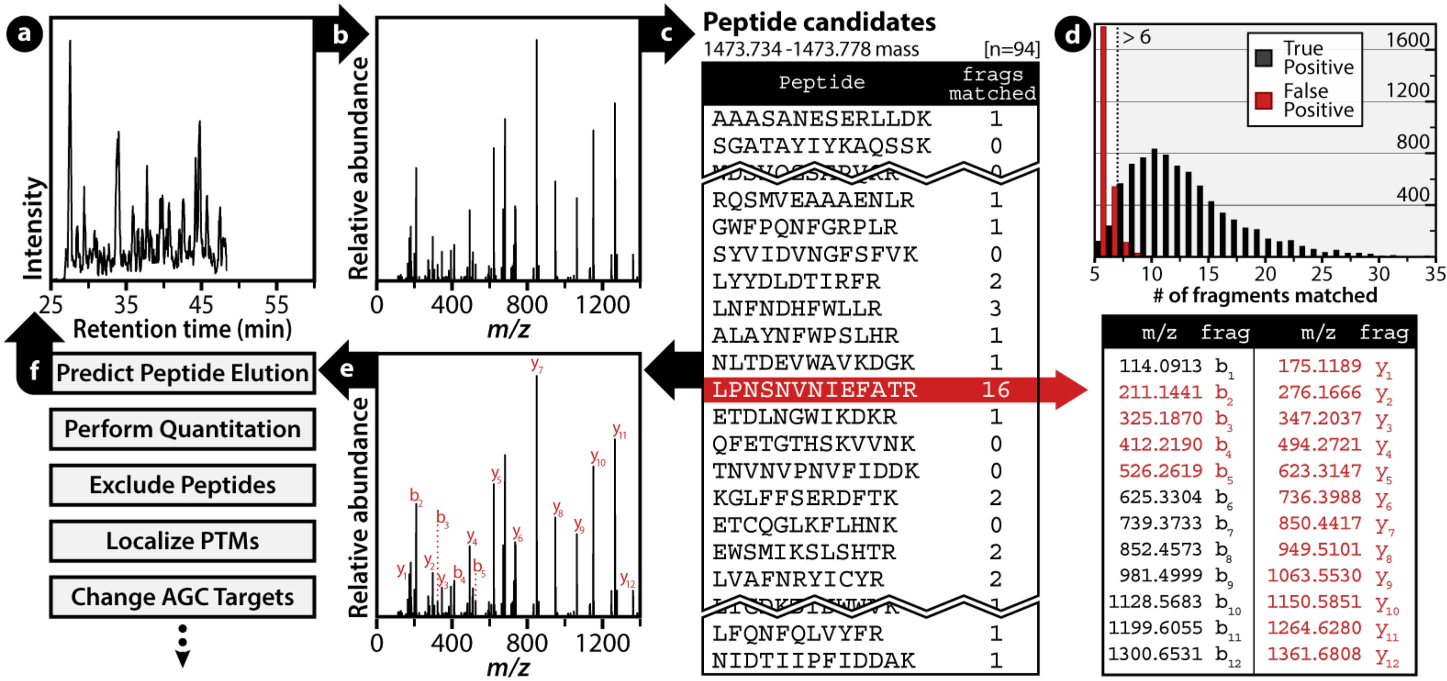
Several DDA add-ons and alternatives have been examined. Sampling depth, for example, can be increased by preventing selection of a *m/z* value identified in a prior technical replicate (PANDA).[9] Irreproducibility can be somewhat countered by informing the DDA algorithm of the precursor *m/z* values of desired targets (inclusion list)—if observed this can ensure their selection for MS/MS. Frequently, however, low abundance peptides may not have precursor signals above noise so that a MS/MS scan, which is requisite for identification, is never triggered. This conundrum is avoided altogether in the data-independent acquisition approach (DIA).[10] Here no attention is paid to precursor abundance, or even presence, instead consecutive *m/z* isolation windows are dissociated and mass analyzed. A main drawback of DIA is that it requires significantly more instrument analysis time as MS/MS scans from every *m/z* window must be collected.[11] As such, DDA analysis remains the preeminent method for MS data acquisition.

Besides improvements in MS analyzer performance, numerous alternative dissociation methods and scan types have recently advanced. These include collision, electron and/or photon-based fragmentation (i.e., trapHCD, HCD, ETD, IRMPD, etc.), specialized quantification scans (i.e., QuantMode), or simply analysis using varied precursor ion targets, *m/z* accuracy, etc.[12–17] Each of these techniques show applicability and superlative performance for a subset of peptide precursors. The result is a dizzying alphabet soup of techniques, scan types, and parameter space that is not easily integrated into the current data acquisition paradigm. Recently we introduced a decision tree (DT) algorithm that used precursor *m*, *z*, and *m/z* to automatically determine, in real-time, whether to employ CAD or ETD during MS/MS.[18] The approach significantly improved sequencing success rates and was an important step in a movement toward development of informed acquisition.

Here we describe the next advance in DT acquisition technology—instant sequence confirmation (*inSeq*). The *inSeq* algorithm processes MS/MS spectra at the moment of collection using the MS system's on-board processing power. With sequence in hand the MS acquisition system can process this knowledge to make autonomous, real-time decisions about what type of scan to trigger next. Here, with the *inSeq* instant identification algorithm, we extend our simple DT method by adding several new decision nodes. These nodes enable novel automated functionalities including: real-time elution prediction, advanced quantification, PTM localization, large-scale targeted proteomics, and increased proteome coverage, among others. This technology provides a direct pathway to transform the current passive data collection paradigm. Specifically, knowing the identity of a peptide that is presently eluting into the MS system permits an ensemble of advanced, automated decision-making logic.

Results and Discussion

Instant sequence confirmation (*inSeq*). To develop an advanced DT acquisition schema, which can seamlessly incorporate the myriad of specialized procedures and scans available on modern day MS systems, we must expedite the spectral analysis process—i.e., from off-line to real-time. There are two obvious pathways to incorporate real-time spectral analysis within an MS system. The first approach exports spectra for processing with an external computing system followed by import of the search outcome.[19] A second, more elegant strategy, is to perform all computation within the MS's on-board computing system.[20] The former approach circumvents complications in accessing instrument firmware and allows for the use of more sophisticated processing power; however, a serious constraint is the time required for import/export of the information (~40 ms). We have pursued technologies and computational algorithms that integrate real-time spectral analysis into the MS system's on-board processors and firmware. We call this method instant sequence confirmation (*inSeq*). Experimental details (e.g., peptide candidates, scan sequences, etc.) are transferred, on-demand, along with the instrument's method file to the instrument before the experiment commences, allowing for flexibility in experimental design with minimum configuration. To establish robustness across platforms we implemented *inSeq* on two distinct MS systems (operating with different code bases)—a dual cell quadrupole linear ion trap-Orbitrap hybrid (LTQ-Velos Orbitrap) and a quadrupole mass filter-Orbitrap hybrid (Q Exactive). In both cases we modified and extended the instrument firmware to quickly (~<20 ms in the case of the more modern Q Exactive system) and accurately (<2% false discovery rate (FDR)) map MS/MS spectra to sequence. The embedded peptide database-matching algorithm processes MS/MS scans immediately (Figure 2.1A&B) by comparing product ions present in the MS/MS scan to those from peptide candidates pre-loaded onto the instrument's firmware. Note the candidate sequences are first filtered so that only sequences whose mass is within a small window (e.g., 5-50 ppm) of the sampled precursor neutral mass are considered (Figure 2.1C). For each candidate sequence the number of +1 product ions (+2 ions are included for precursors >+2) that matched the spectrum at a mass tolerance < 10 ppm is recorded (Figure 2.1C). Next, it uses straightforward scoring metrics, providing



Peptide candidates
1473.734 -1473.778 mass [n=94]

Peptide	frags matched
AAASANESERLLDK	1
SGATAYIYKAQSSK	0
RQSMVEAAAENLR	1
GWFPQNFGRPLR	1
SYVIDVNGFSFVK	0
LYYDLDTIRFR	2
LNFNDHFWLLR	3
ALAYNFWPSLHR	1
NLTDEVWAVKDGK	1
LPNSNVNIEFATR	16
ETDLNGWIKDKR	1
QFETGTHSKVVNK	0
TNVNVPNVFIDDK	0
KGLFFSERDFTK	2
ETCQGLKFLHNK	0
EWSMIKSLSHTR	2
LVAFNRYICYR	2
LFQNFQLVYFR	1
NIDTIIPFIDDAK	1

m/z	frag	m/z	frag
114.0913	b_1	175.1189	y_1
211.1441	b_2	276.1666	y_2
325.1870	b_3	347.2037	y_3
412.2190	b_4	494.2721	y_4
526.2619	b_5	623.3147	y_5
625.3304	b_6	736.3988	y_6
739.3733	b_7	850.4417	y_7
852.4573	b_8	949.5101	y_8
981.4999	b_9	1063.5530	y_9
1128.5683	b_{10}	1150.5851	y_{10}
1199.6055	b_{11}	1264.6280	y_{11}
1300.6531	b_{12}	1361.6808	y_{12}

Figure 2.1: Progression of *inSeq* logic. (A) A nHPLC-MS/MS chromatogram at 48.35 minutes along with an MS/MS scan (B) that was acquired at that time following dissociation of a +2 feature of *m/z* 737.86. Upon collection of the MS/MS scan, *inSeq* groups all peptide candidates (*in silico*, n = 94) whose theoretical mass are within 30 ppm of the experimentally determined precursor neutral mass 1473.756 (C). Then *inSeq* performs *in silico* fragmentation to produce a theoretical product ion series for each of the 94 candidates and proceeds to compare each to the experimental spectrum (<10 ppm mass accuracy). (D) Plot of *inSeq* identifications compared to conventional post-acquisition searching. *inSeq* agrees (True Positive) >98% of the time when >6 fragment ions are matched.

sufficient evidence for the confirmation of a putative sequence without burdening the system with non-essential calculations. On both MS platforms the real-time confirmation algorithm was expediently executed and required no hardware modification, taking an average of 16 ms to perform (Q Exactive, Figure 2.2). To confirm that this small overhead does not affect the overall duty cycle, we compared the number of MS/MS scans performed when *inSeq* was and was not operating (9,076 DDA vs. 8,908 DDA with *inSeq*, ~1.6%). The number of MS scans for the peptide IVGIVSGELNNAAAK within its elution profile further demonstrates the negligible impact on duty cycle as 20 MS scans were taken with *inSeq* inactive as compared to 19 scans with *inSeq* active (Figure 2.3).

To characterize the *inSeq* algorithm we performed a nHPLC-MS/MS experiment on tryptic peptides derived from human embryonic stem cells. A database consisting of all theoretical tryptic peptides (up to three missed cleavages, 6-50 in length) contained within the human proteome was uploaded to the instrument's (Q Exactive) on-board computer. A DDA method was employed and analysis proceeded as usual, except following each MS/MS scan the *inSeq* algorithm was executed and the results logged. This manifest of instant identifications was then compared to those made post-acquisition via traditional database searching at a 1% FDR (target-decoy method). We assumed the conventional post-acquisition approach

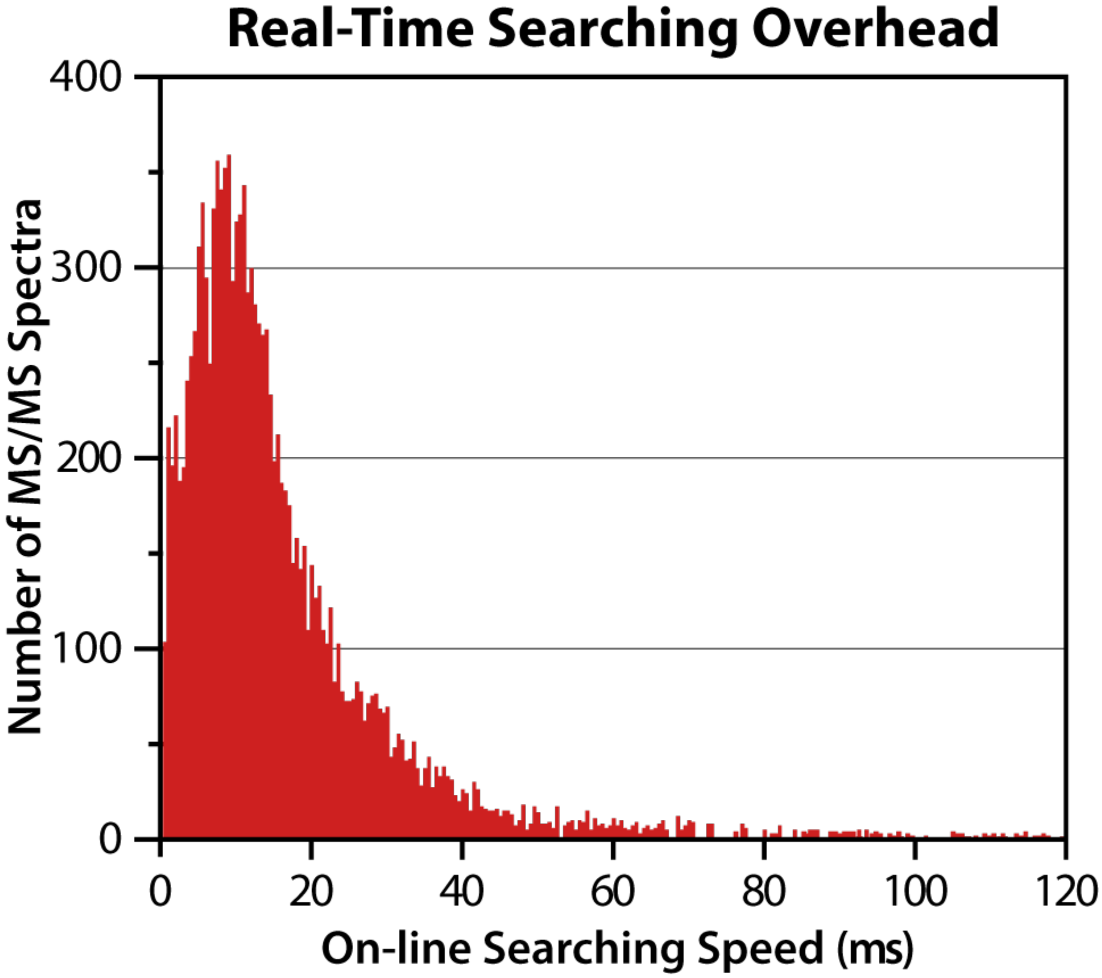


Figure 2.2: Distribution of *inSeq* analysis times (Q Exactive). The complete yeast proteome (6,717 proteins) was digested *in silico*, trypsin specificity up to 3 missed cleavages) and the resulting peptides sorted to only retain those between 6-50 residues in length for a database of 1,174,780 unique sequences. The overhead accrued by *inSeq* is small (μ = 16 ms / spectrum) compared to the overall acquisition rate (~100-250 ms/spectrum), with over 95% of the MS/MS events taking less than 45 ms to search.

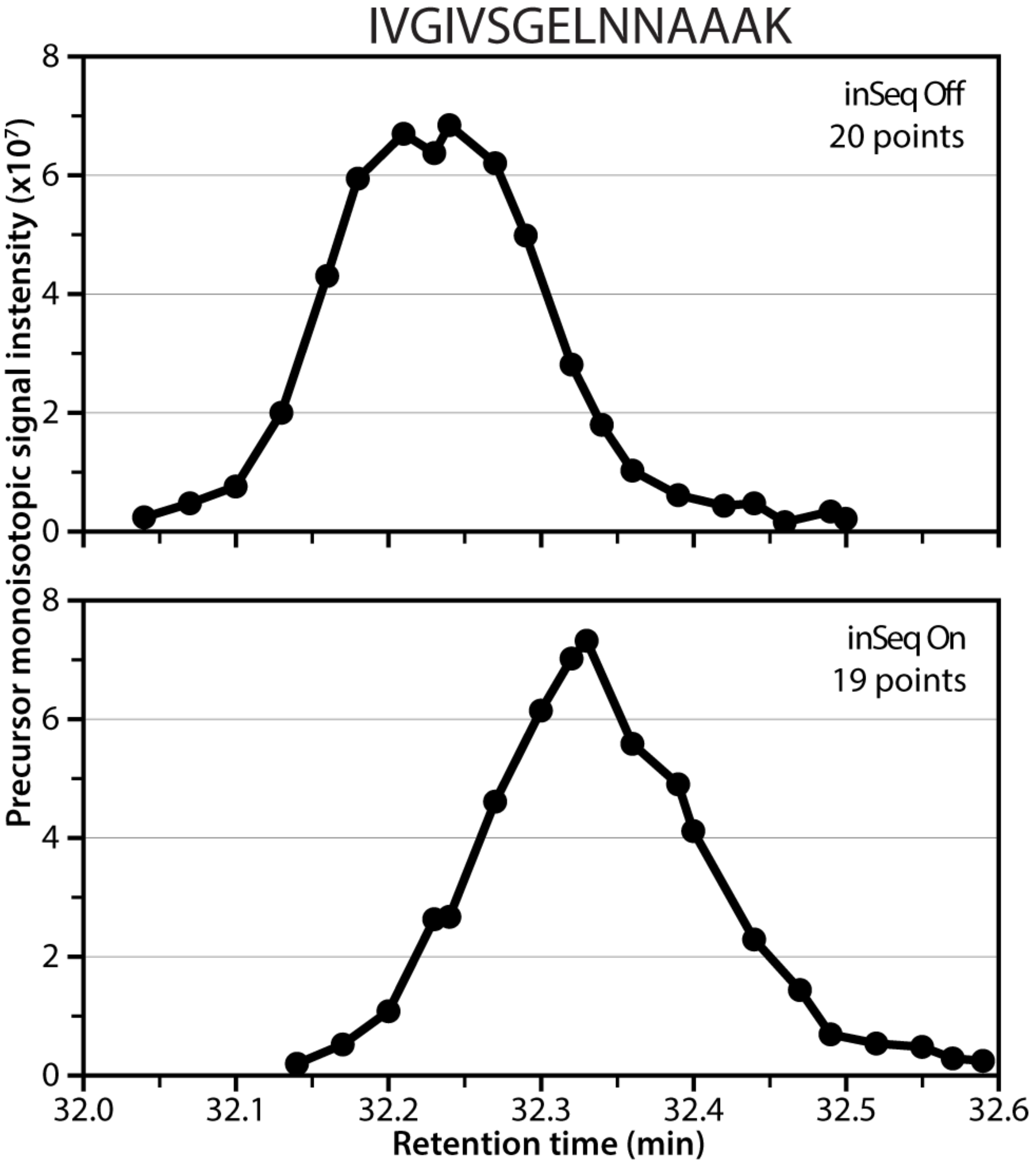


Figure 2.3: Frequency of MS scans with and without *inSeq* active. The extracted ion chromatograms of a random peptide (IVGIVSGELNNAAAK) is displayed for two data-dependent top 10 experiments without (top) and with (bottom) *inSeq* active. In both cases, a large number of MS scans are performed within the elution profile (20 and 19 for *inSeq* off and on, respectively).

to represent the true answer and compared the number of correct instant spectral identifications as a function of matched product ions (Figure 2.1D). From these data we conclude the detection of >6 product ions at high mass accuracy (<10 ppm) by the *inSeq* algorithm produces the correct sequence identification >98% of the time. To determine the impact of *inSeq* on the depth of protein coverage, we compared OMSSA identifications with *inSeq* identifications (species with >6 matching peaks) (Figure 2.4). Traditional post-acquisition searching identified more peptides than *inSeq* (11,095 vs. 7,910, respectively) indicating strong initial performance but also room for further development of a more sophisticated real-time scoring algorithm.

inSeq represents a straightforward approach to correlate sequence to spectrum and is positioned to become an essential technology in transforming the current passive data collection paradigm. Specifically, learning the identity of a peptide that is presently eluting into the MS system permits an ensemble of advanced, automated decision-making logic. These concepts build upon our previous development of the data-dependent decision tree (DT) method. There we embedded an on-board algorithm to make unsupervised, real-time decisions of which fragmentation method to engage, based on precursor charge (*z*) and *m/z*. Here, with the *inSeq* instant identification algorithm, we extend our simple DT method by adding several new decision nodes (Figure 2.5). These nodes enable automated functionalities

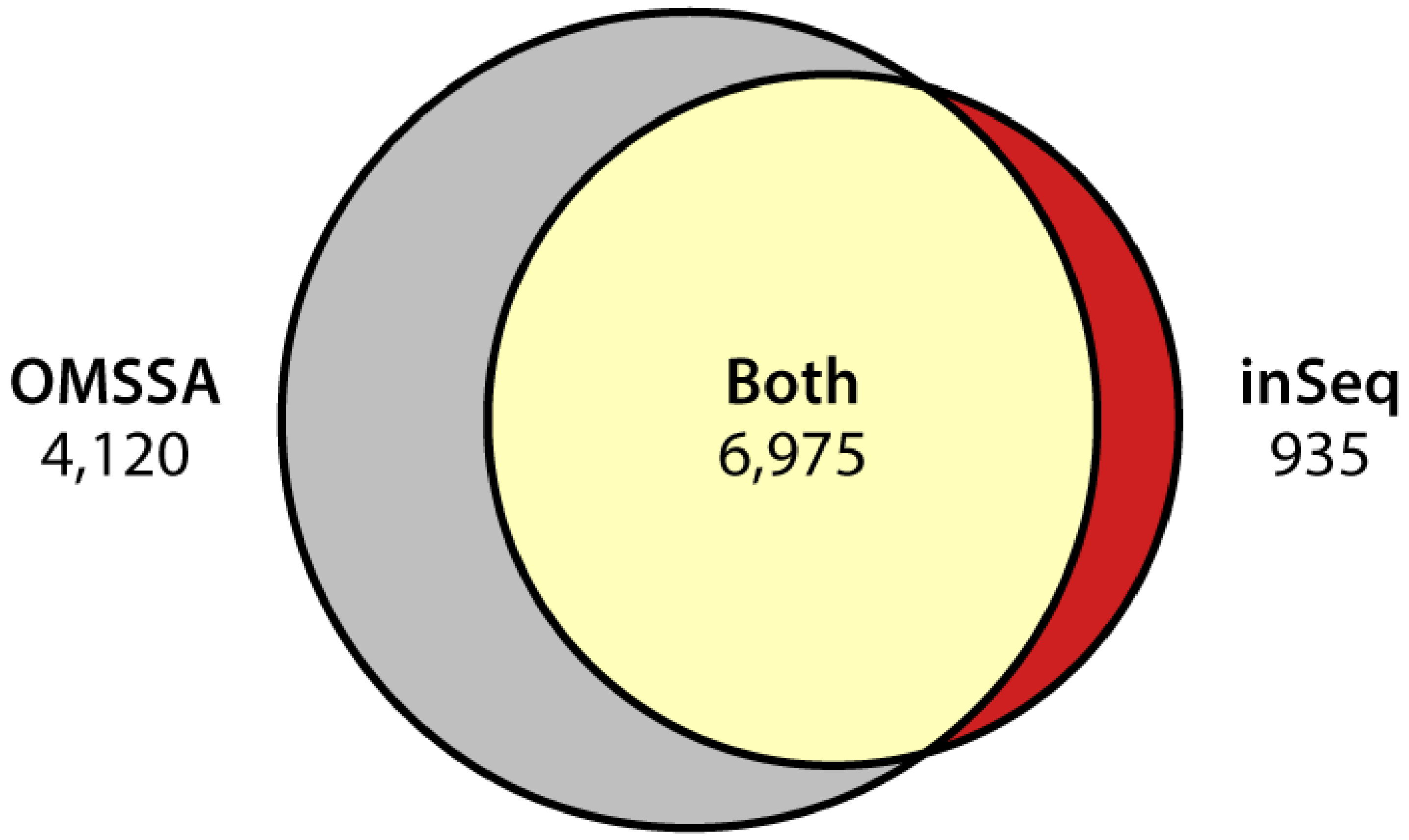


Figure 2.4: Overlap of *inSeq* and OMSSA identifications. Each spectrum in a 120 min nHPLC-MS/MS gradient was scored with OMSSA and *inSeq*. OMSSA identifications were filtered to 1% FDR and *inSeq* identifications were confirmed only if the spectra contained >6 matching peaks, as described in the text. Overlap between OMSSA and *inSeq* demonstrates that *inSeq* identifications are usually correct. The OMSSA-only identifications represent a subset of spectra which are false negatives with respect to *inSeq*.

including: real-time elution prediction, advanced quantification, PTM localization, large-scale targeted proteomics, and increased proteome coverage, among others (Figure 2.1F).

Predicting peptide elution. Liquid chromatography is the conventional approach to fractionate highly complex peptide mixtures prior to measurement by MS. The highest MS sensitivity is achieved when one tunes the MS system to detect a given target (i.e., execute MS/MS) regardless of its presence in the preceding MS event (i.e., selected reaction monitoring). SRM measurements deliver both sensitivity and reproducibility at the cost of bandwidth. Specifically, if one does not know the elution time of a target, the duration of the nHPLC-MS/MS analysis must be dedicated to conditions for that specific entity. If elution times are known, then multiple SRM scan events can be programmed allowing for detection of multiple targets; however, chromatographic conditions must remain identical or the scheduled SRM elution windows will no longer align. Still, the bandwidth of that approach is low (~100 peptide targets per nHPLC-MS/MS analysis) and compiling such an experiment is highly laborious.[21]

We surmised that *inSeq* could inform the MS system, without human intervention, of which peptide targets are most likely to subsequently elute. Such capability could enable robust, large-scale targeting (>500 per analysis) in an au-

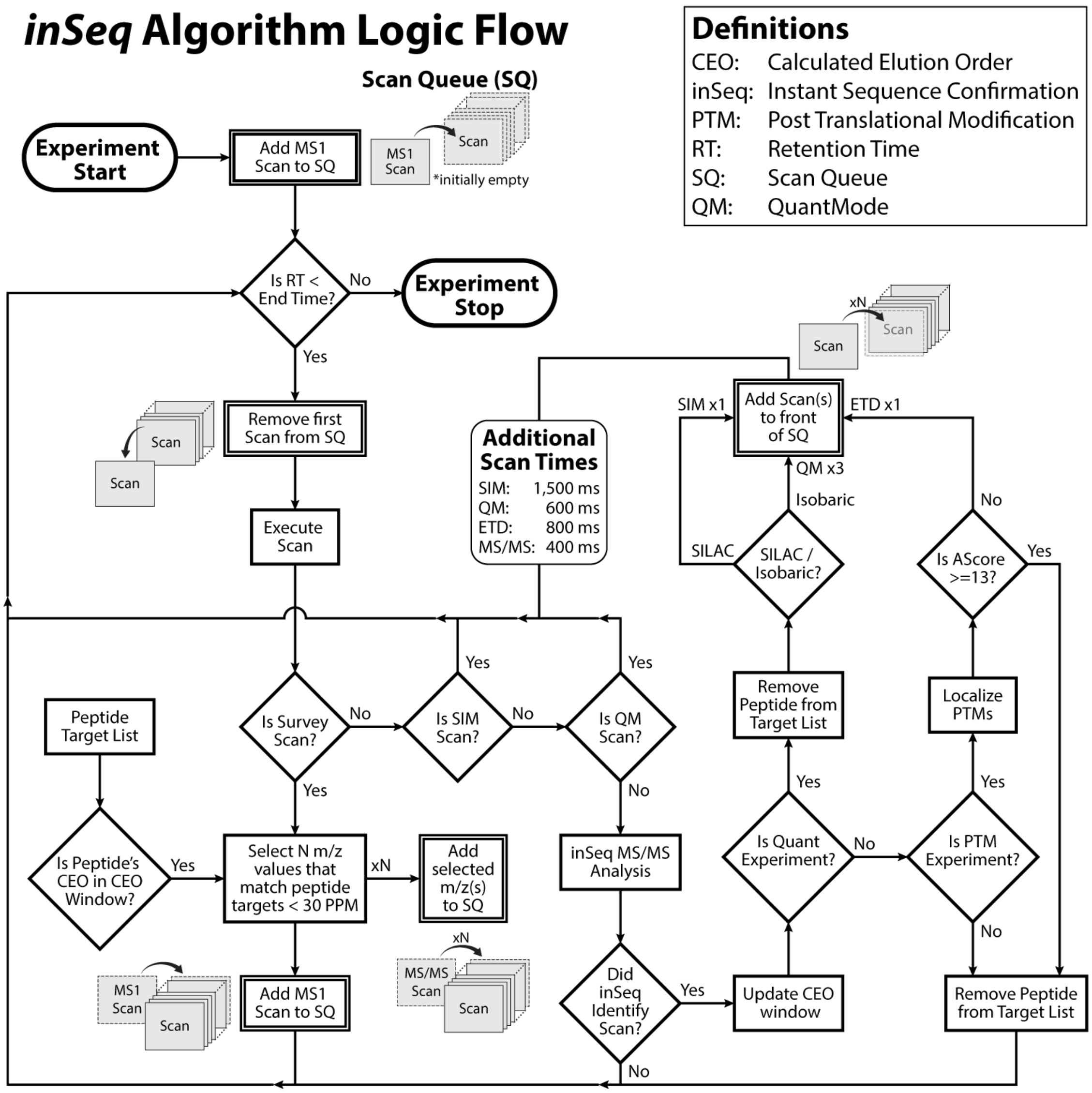
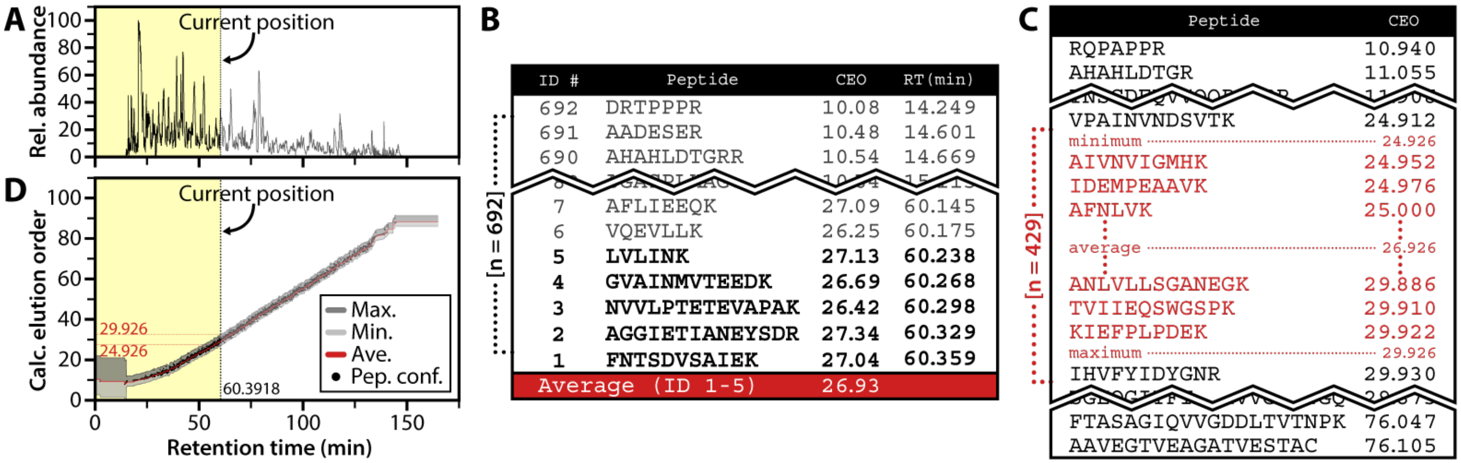


Figure 2.5: Basic information flow for implementing *inSeq*. The flow of *inSeq* follows a targeted Top-N inclusion list (Target List) routine with additional analysis steps interspersed. Following an MS scan, the top 10 peaks that match precursors in the Target List (filtered by elution order) are added to a Scan Queue (SQ). Each MS/MS scan is analyzed with *inSeq* to determine if the peptide of interest is there. If identified, the peptide is removed from the Target List and additional analyses for quantitation and PTM localization may be performed.

tomated manner. Our approach relies upon relative peptide elution order and, consequently, bypasses the use of absolute retention times, which shift depending on chromatographic conditions and are not directly portable from multiple disparate experiments. Peptide elution order can be obtained in two ways: First, discovery experiments can be employed to determine retention order by normalizing the measured retention time for each detected peptide sequence. Second, the relative hydrophobicity for any sequence can be theoretically determined using existing software (e.g., SSRCalc).[22–24] In our experience experimentally determined retention order offers better precision; still, it requires prior knowledge which may not be available. However retention order is determined, the real-time confirmation algorithm maintains a rolling average of the calculated elution order (CEO—a number describing the relative elution order of a target peptide) so that target peptides having nearby CEOs are specifically pursued (Figure 2.1B). Figure 2.6 presents an overview of this approach. This example, 60.39 minutes into the chromatograph, highlights the last five *inSeq* identified peptides and their average CEO (26.926 a.u.). The on-board algorithm then computes an asymmetric CEO window (5 a.u., 24.926-29.926) that presents a short list of desired targets having CEOs within that range (Figure 2.1C). With this information the MS system can trigger specialized MS/MS scans specific to this refined target subset. Note that as targets are identified, the



ID #	Peptide	CEO	RT (min)
692	DRTPPPR	10.08	14.249
691	AADESER	10.48	14.601
690	AHAHLDTGRR	10.54	14.669
7	AFLIEEQK	27.09	60.145
6	VQEVLLK	26.25	60.175
5	**LVLINK**	**27.13**	**60.238**
4	**GVAINMVTEEDK**	**26.69**	**60.268**
3	**NVVLPTETEVAPAK**	**26.42**	**60.298**
2	**AGGIETIANEYSDR**	**27.34**	**60.329**
1	**FNTSDVSAIEK**	**27.04**	**60.359**
Average (ID 1-5)		26.93	

Peptide	CEO
RQPAPPR	10.940
AHAHLDTGR	11.055
VPAINVNDSVTK	24.912
AIVNVIGMHK	24.952
IDEMPEAAVK	24.976
AFNLVK	25.000
ANLVLLSGANEGK	29.886
TVIIEQSWGSPK	29.910
KIEFPLPDEK	29.922
IHVFYIDYGNR	29.930
FTASAGIQVVGDDLTVTNPK	76.047
AAVEGTVEAGATVESTAC	76.105

Figure 2.6: Elution order prediction using *inSeq*. (A) Experimental chromatogram, 60.39 minutes into a 120 minute gradient, obtained while *inSeq* was recording instant identifications. (B) The *inSeq* calculated elution order (CEO) obtained by averaging the CEOs of the preceding 5 instantly identified peptides. (C) The asymmetric window (5 a.u.) surrounding the instant CEO average (μ = 26.926) and their corresponding sequences. (D) This analysis is repeated following each new peptide confirmation to constantly realign the CEO window based on current chromatographic conditions.

CEO window is dynamically adjusted so targets come into and out of the range precisely when they are eluting.

To test this technology, we performed a DDA nHPLC-MS/MS experiment in which tryptic peptides from a human ES cell sample were separated over a 60 minute gradient. Following data collection the resulting MS/MS spectra were mapped to sequence using database searching (1% FDR). The unique peptide identifications (4,237) were sorted by observed retention time—this ordering then served as the CEO. 3,000 of these peptides were randomly selected as ”targets” and loaded onto the instrument firmware (Velos-Orbitrap), along with their respective CEO, as a database for *inSeq*. The sample was then re-analyzed with *inSeq* activated, but with a doubled gradient length (120 min). Figure 2.6D displays the CEO window as calculated in real-time by the MS system (*inSeq*) plotted beside the actual elution time of identified peptides. Greater than 95% of the peptides (2,889) fell within the rolling CEO window and were identified by both *inSeq* and post-acquisition searching. At our present capability we can achieve window widths similar to those used in absolute scheduling type experiments (~3-6 minutes) on a scale that is 30X larger (e.g., 3,000 targets vs. 100) with minimal effort.[21,25] Further, we demonstrate that our approach adapts to different chromatographic conditions with no negative effects (Figure 2.6D). The key to the high portability and simplicity of our algorithm is the use of *inSeq* for continual, real-time realignment.

Improvement of quantitative accuracy. The method of stable isotope labeling has greatly propelled large-scale, quantitative analysis.[26–31] While generally robust, these techniques can yield spotty data for certain peptide and protein groups—mainly those present at low abundances. For SILAC, low signal-to-noise (S/N) precursor peaks in the MS scan often result in either omission of that particular feature or quantitative imprecision, if included.[32] For isobaric tagging, low intensity reporter ion signals (MS/MS) induce similar shortcomings.[33] We surmised that *inSeq* could be employed to counter these limitations.

First, we developed an *inSeq* module to improve the quality of isobaric label-based measurements. The module analyzes MS/MS spectra, using *inSeq*, and, when a peptide of interest is detected, the quality of quantitative data is assessed. Should the reporter ion signals fall below a specified threshold, *inSeq* triggers follow-up scans to generate increased signal at the very instant the target peptide is eluting. In one implementation, we instructed *inSeq* to automatically trigger three quantitative scans, using the recently developed QuantMode (QM) method, to generate superior quality quantitative data on targets of high value.[16] The trio of QM scans are then summed offline.

To assess this decision node we analyzed a sample comprising three biological replicates of human embryonic stem cells pre- and two days-post bone morphogenetic protein 4 (BMP4) treatment (i.e., TMT 6-plex, three pre-treatment and three post BMP4 treatment cell populations). BMP4, a growth factor that induces context-dependent differentiation in pluripotent stem cells, is widely used to study differentiation to biologically relevant cell lineages such as mesoderm and endoderm.[34–36] Whenever a target peptide was identified by *inSeq*, three QM scans were immediately executed. This ensured that all identified peptides had the same number of quantitation scans, enabling a direct comparison for analyzing multiple QM scans within this experiment. Figure 2.7A demonstrates the benefit of summing isobaric tag intensities from one, two, or three consecutive quantitation scans for an *inSeq* identified target peptide having the sequence FCADHPFLFFIR from the protein SERPINB8. Here the ratio of change between control and treatment cell lines measured in one QM scan is large (5.86) but not significant (P = 0.067, Student's t-test with Storey correction).[37] Note significance testing was accomplished by assessing variation within the three biological replicates of both treatment and control cell lines. The measured ratio remains relatively unchanged (5.22 and 5.43) as reporter tag signals from additional quantitation scans are added; however, the corresponding P-values decrease to 0.014 and 0.012 when two or three quantitation scans are summed. By plotting the $\log_2$ ratio of quantified proteins from the three

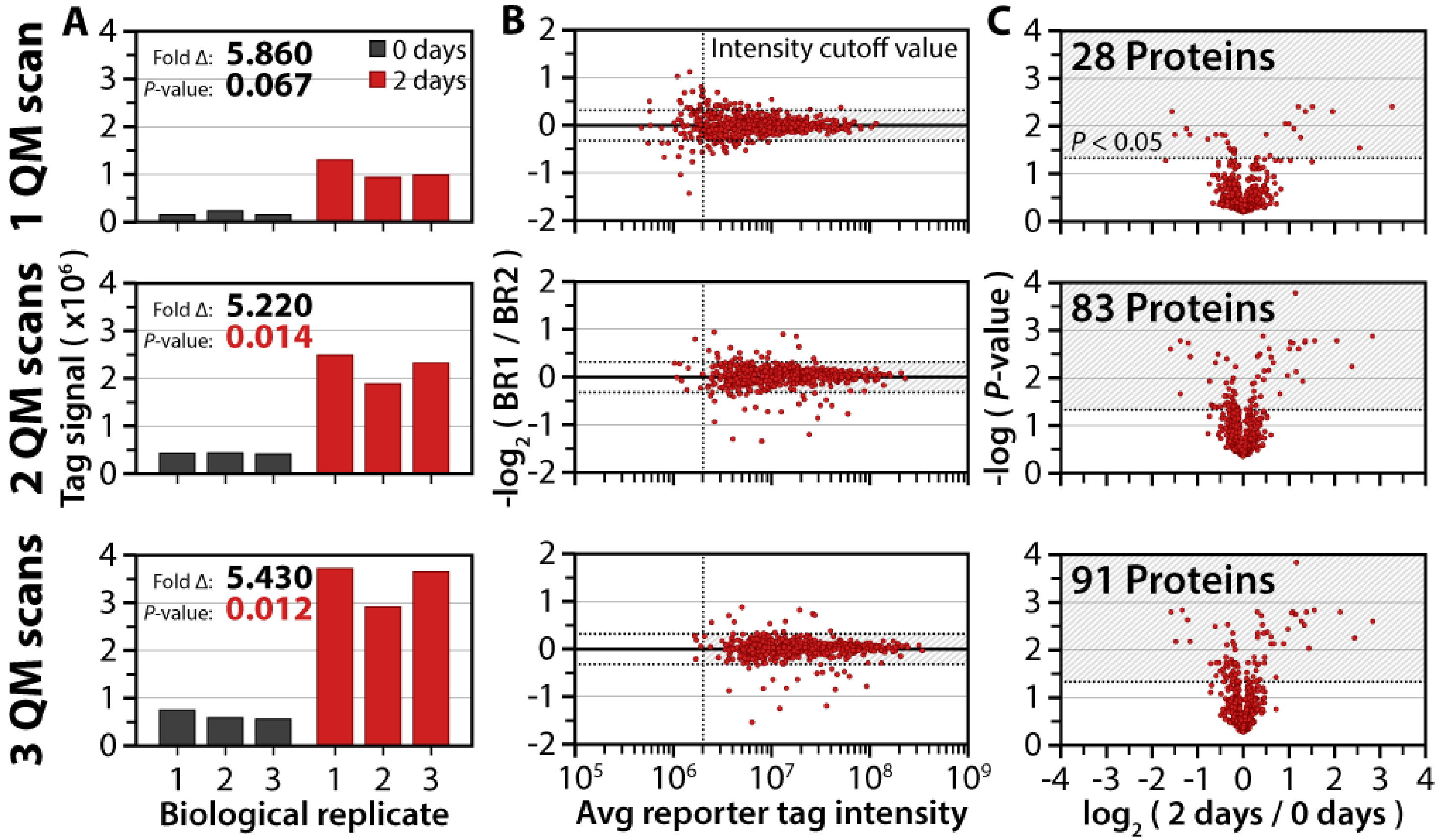


Figure 2.7: *inSeq* improves quantitative outcomes for isobaric tagging. An *inSeq* decision node was written so that a real-time identification of a target sequence prompted automatic acquisition of three consecutive QuantMode (QM) scans. (A) Summing the reporter ion tag intensity from one, two, or three QM scans greatly improves the statistical significance of the measurement. (B) Summation of QM technical replicates reduces the variation in biological replicate measurement by increasing reporter ion S/N. (C) *inSeq*-triggered QM scans increase the number of significantly changing proteins from 28 to 91.

biological replicates against the average intensity of isobaric labels (Figure 2.7 B) we demonstrate this improved significance results from boosted reporter S/N. Ideally this $\log_2$ ratio would be zero, indicating perfect biological replication; however, when only one quantitation scan is employed this ratio severely deviates from zero with decreasing tag intensity. To improve overall data quality and to omit potentially erroneous measurements we, and others, employ arbitrary reporter signal cutoffs (dashed vertical line in Figure 2.7B). Summation of additional quantitation scans increases the average reporter tag intensity, raising nearly all of the protein measurements above the intensity cutoff value (74, 9, and 4 proteins omitted using one, two, and three quantification scans, respectively). This quantification decision node also increased the number of proteins within 25% of perfect biological replication (horizontal dashed line).

To determine if the method could improve the number of statistically significant differences between the cell populations, we calculated the $\log_2$ ratio of treated vs. control (i.e., 2 days/0 days) for each of the 596 quantified proteins ($P<0.05$, Student's t-test with Storey correction, Figure 2.7C). Only 28 proteins display significant change when one QM scan is used. By simply adding the reporter tag signal from additional scans the number of significantly changing proteins increases nearly threefold, from 28 to 91 when all three QM scans are analyzed together. Many stable isotope incorporation techniques measure heavy and light peptide pairs in MS (e.g., SILAC). This approach, of course, requires the detection of both partners; note low abundance peptides are often identified with low, or no, precursor signal in the MS. We supposed that addition of another *inSeq* decision node could circumvent this problem. We cultured human embryonic stem cells in light and heavy media. Protein extract from these cultures was mixed 5:1 (light:heavy), before digestion overnight with LysC. The SILAC node was developed to select precursors from an MS scan only if the monoisotopic mass was within 30 ppm of any target on a list which contained 4,000 heavy and light peptides from a previous discovery run. Targets were selected only if the SILAC ratio deviated from the expected ratio of 5 by 25%, i.e., the subset containing the most error. Following MS/MS, the resulting spectra were analyzed using *inSeq*. When a target of interest was identified, *inSeq* instructed the system to immediately record a SIM scan surrounding the light/heavy pair with a small, charge-dependent isolation window (~8-10 Th).

The average ratio of the light and heavy peptides subtly, but significantly, shifted from 4.47 under normal analysis to 5.34 for the *inSeq* triggered SIM scans (Student's t-test, p-value < 6×10^{-20}). More importantly, the number of useable measurements, i.e., when both partners of the pair are observed, increased by ~20% (2,887 under normal analysis to 3,548 with *inSeq*, Figure 2.8A). Figure 2.8B displays an example

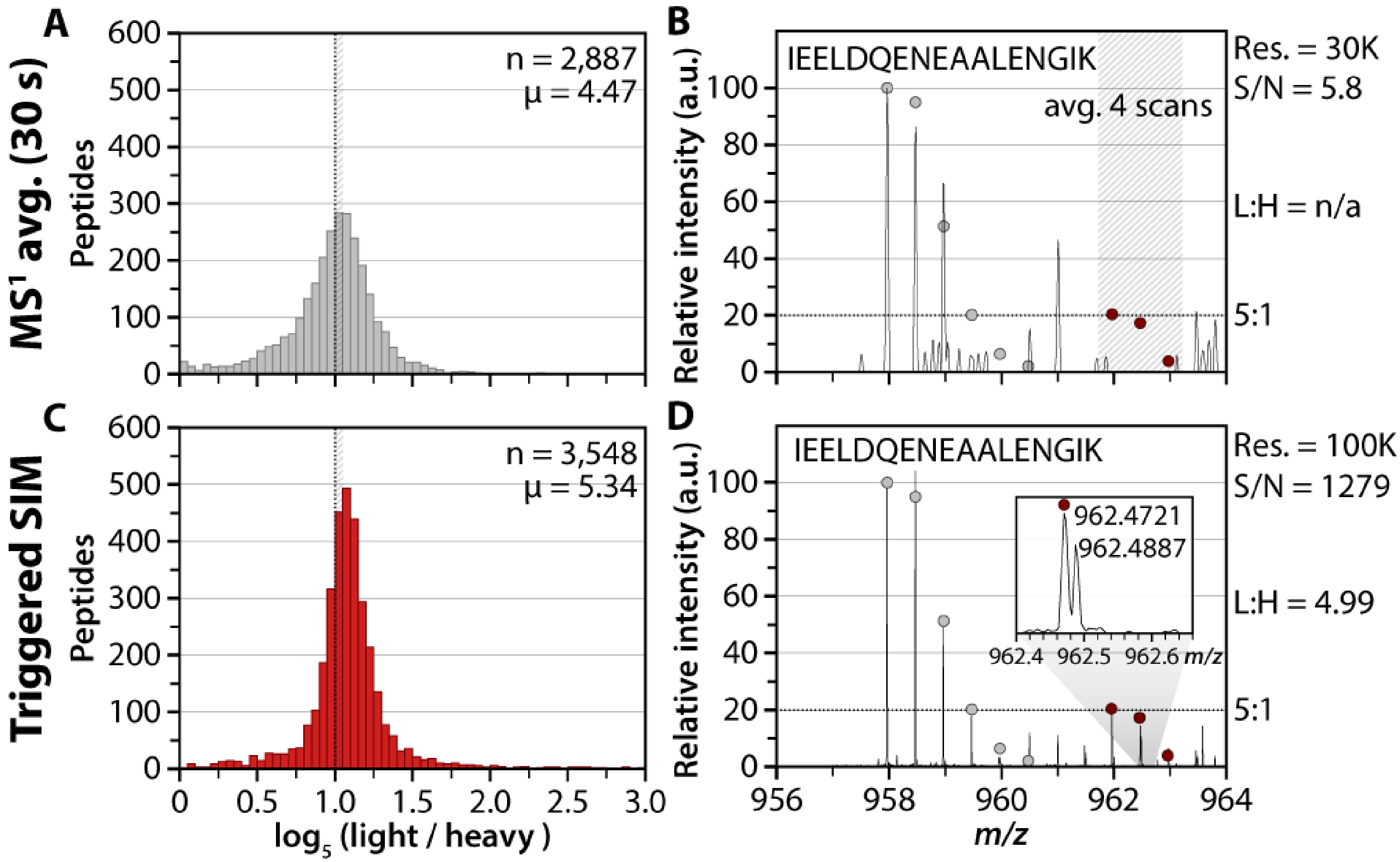


Figure 2.8: ***inSeq* can improve quantitative outcomes for SILAC.** Following an *inSeq* confirmation of a peptide having the sequence, IEELDQENEAALENGIK, a narrow (8 Th), high-resolution (R = 100,000) SIM scan was automatically triggered and increased the S/N from 5.8 to 1,279 (Panels A and B). This SIM scan enabled detection of both partners and yielded the correct ratio of 5:1 (light:heavy). Besides increased dynamic range, the theoretical isotope distribution (shown in open circles) closely matches in the SIM scan (B), while the signal for the heavy partner is not even detectable in the MS (A). Over our entire data set, the *inSeq* triggered SIM scans improved the mean ratio from 4.47 to 5.34, but, more impressively, produced ~20% more quantifiable measurements (3,548 vs. 2,887).

of the *inSeq*-triggered SIM scan and the increase in S/N and accuracy it affords. Here the MS/MS scan of the light partner was mapped, in real-time, to the sequence IEELDQENEAALENGIK. This event triggered a high resolution SIM scan (8 Th window), which led to the ratio of 4.99:1 (correct ratio 5:1). Here gas phase enrichment was essential to quantify the relative abundance, as the isotopic envelope of the heavy partner was not observed, even with extensive spectral averaging of successive MS scans (~30 s, Figure 2.8B). Whether for MS or MS/MS centric methods, we conclude that *inSeq* technology will significantly improve the quality of quantitative data with only a minimal impact on duty cycle.

Post-Translational Modification Site Localization. The presence of post-translational modifications (PTMs) on proteins plays a major role in cellular function and signaling. Unambiguous localization of PTMs to residue demands observation of product ions resulting from cleavage of the residues adjacent to the site of modification, i.e., site-determining fragments (SDFs). In a typical analysis only about half of the identified phosphorylation sites can be mapped with single amino acid resolution, stymying systems-level data analysis. We reasoned that *inSeq* could be leveraged to boost PTM localization rates by dynamically modifying MS/MS acquisition conditions when necessary. As such we developed an online PTM localization decision node to determine, within milliseconds, whether a MS/MS spectrum contains SDFs to unambiguously localize the PTM. Should SDFs be lacking, *inSeq* instantly orchestrates further interrogation.

The PTM localization node is engaged when *inSeq* confirms the detection of a PTM-bearing peptide. After the sequence is confirmed, *inSeq* assesses the confidence with which the PTM(s) can be localized to a particular amino acid residue. This procedure is accomplished by computing an online probability score similar to post-acquisition PTM localization software—i.e., AScore.[38] Briefly, *inSeq* compares all possible peptide isoforms against the MS/MS spectrum. For each SDF the number of matches at <10 ppm tolerance is counted and an AScore is calculated (*inSeq* uses similar math). If the AScore of the best fitting isoform is above 13 ($p < 0.05$) the PTM is declared localized. When the AScore is below 13, however, *inSeq* triggers further characterization of the eluting precursor until either the site has been deemed localized or all decision nodes have been exhausted. Additional characterization can include many procedures such as acquisition of MS/MS spectra using different fragmentation methods (e.g., CAD, HCD, ETD, PD, etc.), varied fragmentation conditions (e.g., collision energy, reaction time, laser fluence, etc.), increased spectral averaging, MSn, pseudo MSn, modified dynamic exclusion, and altered AGC target values, among others.[39]

To obtain proof-of-concept results we wrote a simple *inSeq* node that triggered

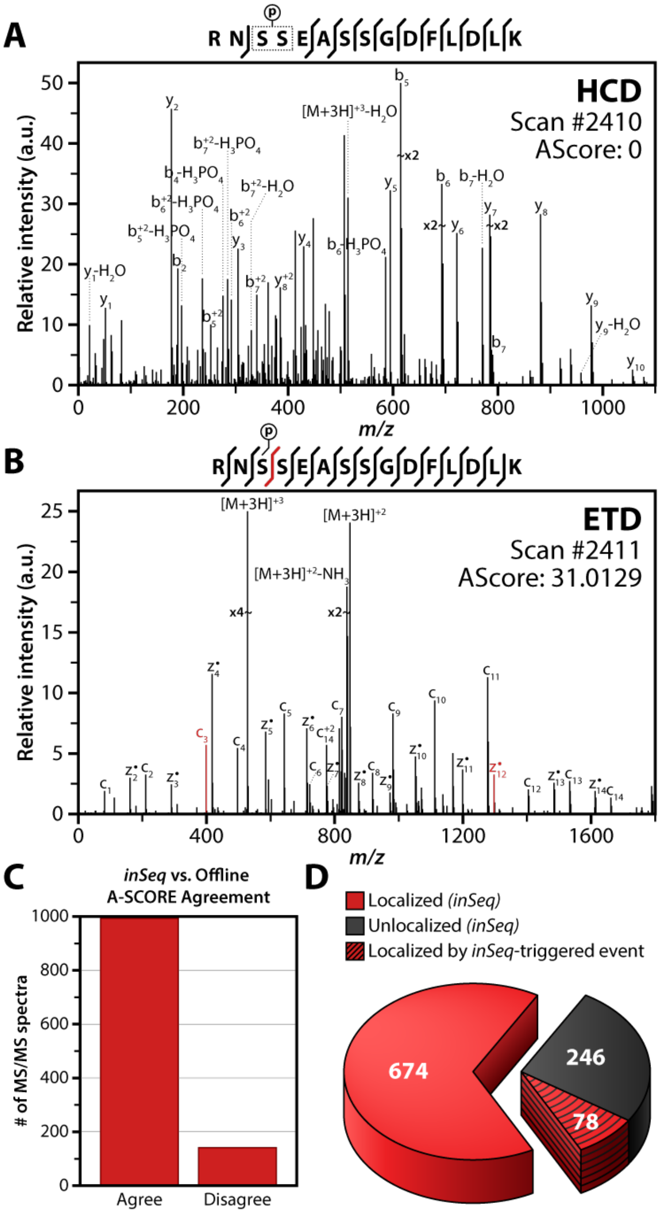


Figure 2.9: *inSeq* can improve PTM localization rates. Following MS/MS (HCD) of the singly-phosphorylated precursor RNsSEASSGDFLDLK, *inSeq* could not find sufficient information to confidently localize the modification to either Ser 3 or 4 (A, AScore = 0). *inSeq* immediately triggered an ETD MS/MS scan event on the same precursor (B). This spectrum was assigned an AScore of 31.0129 (phosphorylation on Ser3) and was considered confidently localized—note the SDFs c_3 and $z\cdot_{12}$ ions. (C) Globally, the *inSeq* localization calculation agreed with offline analysis using the actual AScore algorithm. (D) Using a simple dissociation method DT, *inSeq* produced a confidently localized phosphorylation site for 78 of 324 unlocalizable sites, saving nearly 25% of them.

an ETD MS/MS scan of phosphopeptides that were not localized following HCD MS/MS. In one example (Figure 2.9) the sequence, RNSSEASSGDFLDLK, was confirmed to contain a phosphoryl group; however, the *inSeq* algorithm could not confidently localize the PTM to any of the four Ser residues (AScore = 0). Next, *inSeq* triggered an ETD MS/MS scan of the same precursor (Figure 2.9B). The resulting spectrum was then analyzed for the presence of the SDFs, c_3 / $z\cdot_{12}$. Both of these fragments were present and the site was localized to Ser 3 with an AScore of 31.0129 ($p < 0.00079$). Post-acquisition analysis confirmed the results of our online *inSeq* approach—both spectra (HCD and ETD) were confidently identified and their calculated AScores were 0 and 45.58, respectively. When compared on a global scale, 993 of the 1,134 *inSeq*-identified phosphopeptides had localization judgments that matched post-acquisition AScore analysis (Figure 2.9). This slight difference is the result of using different localization algorithms for online and post-acquisition analysis. Primarily, the post-acquisition method considers fragment ions on either side of the site-determining fragments separately, while the *inSeq* method does perform this extra step for simplicity.[30,38] These data demonstrate that our localization node is highly effective at instantaneously determining whether a PTM site can be localized. Unfortunately, only marginal gains were achieved in this basic implementation as most precursors were doubly charged and, therefore, not effectively sequenced by ETD. Next, we modified the *inSeq* decision node to incorporate a dissociation method DT. Here a follow-up ETD or combination ion trap CAD/HCD scan was triggered depending upon precursor charge (*z*) and *m/z*. With the slightly evolved algorithm the *inSeq* method detected 998 phosphopeptides in a single shotgun experiment. It determined that 324 of these identifications lacked the information to localize the PTM site and, in those cases, triggered the new dissociation decision node. 78 of these unlocalizable sites were confidently mapped with this technique—salvaging nearly 25% of the unlocalized sites (Figure 2.9D). These encouraging results demonstrate that *inSeq* has great promise to curtail the problem of PTM localization in a highly automated fashion. We note there are dozens of parameters to explore in the continued advancement of this PTM localization decision node.

Conclusion

Here we described an instant sequencing algorithm (*inSeq*) that operates using the pre-existing processors of the MS. Rapid real-time sequencing affords several novel data acquisition opportunities. To orchestrate these opportunities we constructed an advanced decision tree logic that extends our earlier use of the method to intelligently select dissociation type. The approach can circumvent longstanding problems with the conventional DDA paradigm. We provided three such examples herein. First, we demonstrated that knowledge of which peptide sequences are eluting can facilitate the prediction of soon to elute targets. This method shows strong promise to revolutionize the way in which targeted proteomics is conducted. Second, we used quantitative decision nodes that fired when *inSeq* detected a peptide sequence of interest. With either SILAC or isobaric tagging, significant gains in quantitative outcomes were documented. Third, we endowed *inSeq* with an instant PTM site-localization algorithm to determine whether or not to initiate more rigorous follow-up at the very instant the peptide of interest was eluting. We show that the *inSeq* site localizer is highly effective (90% agreement with post-acquisition analysis) and that triggering a simple dissociation method DT can improve site localization by ~25%. Further development will doubtless deliver additional gains.

Targeted proteomics is an area of increasing importance. Following discovery analysis it is natural to cull the list of several thousand detected proteins to several hundred key players. In an ideal world these key proteins are then monitored in dozens or even hundreds of samples with high sensitivity and reproducibility, without rigorous method development and be expediently performed. We envision that advanced DT analysis with *inSeq* could offer such a platform. Using the retention time prediction algorithm we introduced here one can foresee the *inSeq* algorithm quickly and precisely monitoring hundreds of peptides without the extensive labor and pre-planning required by the selected reaction monitoring (SRM) technique, current state-of-the-art.[40] Other possibilities include automated pathway analysis where user-defined proteins, within a collection of pathways, are simply uploaded to the MS system. Then, *inSeq* automatically determines the best peptides to track, their retention times, and constructs the method. Two key advantages over current SRM technology make this operation possible. First, knowledge of specific fragmentation transitions are not necessary as all products are monitored with high mass accuracy. Second, precise elution time scheduling is not necessary as *inSeq* can use CEO, experimental or theoretical, to dynamically adjust the predicted elution of targets. In this fashion the most tedious components of the SRM workflow can be avoided.

Experimental Methods

Cell Culture. Human embryonic stem cells (line H1) were maintained in feeder independent media as previously described.[41] For SILAC experiments, DMEM/F12 lacking lysine and arginine (Mediatech Inc.) was supplemented with light arginine (Sigma-Alrich) and either heavy labeled lysine (Cambridge Isotopes Laboratories) or light lysine (Sigma-Aldrich). Cells were cultured on Matrigel (BD Biosciences) and split 1:8 at approximately 80% confluency using 0.1 mM EDTA. To harvest cells, TripLE Express (Invitrogen) was applied for five minutes at 37°C. Following cell detachment, an equivalent volume of ice-cold DPBS (Invitrogen) was added before centrifugation. Cell pellets were subsequently washed twice in ice-cold DPBS and stored at -80°C. BMP4-treated cells were grown and harvested as described above, except that 5 ng/mL BMP4 (R&D Systems) was added into the media and cells were split using TrypLE (Invitrogen). For BMP4 experiments, single cells were plated at the density of $4{\times}10^4/\text{cm}^2$, for 2 days of treatment. We collected ~10^8 cells for each analysis.

Cell Lysis. For all analysis, human embryonic stem cells were lysed in ice-cold 8 M urea, 40 mM NaCl, 50 mM tris (pH 8), 2 mM $MgCl_2$, 50 mM NaF, 50 mM β-glycerol phosphate, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, 1X mini EDTA-free protease inhibitor (Roche Diagnostics), and 1X phosSTOP phosphatase inhibitor (Roche Diagnostics). To solubilize protein and ensure complete lysis, samples were sonicated three times for 15 seconds with 30 second pauses. Total protein was then quantified using a BCA protein assay kit (Thermo Scientific Pierce).

Isobaric Label Sample Preparation. For analysis, 250 μg of protein from each sample was reduced by adding DTT to a final concentration of 5 mM, and alkylated with 15 mM iodoacetamide before final capping with 5 mM DTT. Digestion was carried out by adding LysC (Wako Chemicals) at a 1:100 enzyme-to-protein ratio and incubating at 37°C for 2 hours. At this time, the lysate was diluted with 25 mM tris (pH 8) to a final urea concentration of 1.5 M and further digested for 12 hours at 37°C with trypsin (Promega) at a 1:100 enzyme to protein ratio. Peptides were then acidified with TFA to quench the reaction and de-salted using C-18 solid phase extraction (SPE) columns (Waters). TMT labeling was carried out per manufacturer's directions (Thermo Scientific Pierce). Samples were mixed in a 1:1:1:1:1:1 ratio before analysis.

SILAC Sample Preparation. Protein from the light and heavy embryonic stem cell cultures was mixed in a 5:1 ratio (light:heavy) by pooling 2.5 mg of light protein and 0.5 mg of heavy protein. The sample was reduced by adding DTT to a final concentration of 5 mM, and alkylated with 15 mM iodoacetamide before final capping with 5 mM DTT. Digestion was carried out by adding LysC (Wako Chemicals) at a 1:100 enzyme-to-protein ratio and incubating at 37°C overnight. Peptides were then acidified with TFA to quench the reaction and de-salted using C-18 solid phase extraction (SPE) columns (Waters).

Phosphopeptide Sample Preparation. From an embryonic stem cell culture, 1 mg of protein was reduced by adding DTT to a final concentration of 5 mM, and alkylated with 15 mM iodoacetamide before final capping with 5 mM DTT. Digestion was carried out by adding LysC (Wako Chemicals) at a 1:100 enzyme-to-protein ratio and incubating at 37°C for 2 hours. At this time, the lysate was diluted with 25 mM tris (pH 8) to a final urea concentration of 1.5 M and further digested for 12 hours at 37°C with trypsin (Promega) at a 1:100 enzyme to protein ratio. Peptides were then acidified with TFA to quench the reaction and de-salted using C-18 solid phase extraction (SPE) columns (Waters). Phosphopeptides were enriched via immobilized metal affinity chromatography (IMAC) using magnetic beads (Qiagen). Following equilibration with water, the beads were treated with 40 mM EDTA (pH 8.0) for 30 minutes with shaking, and washed 3X with water again. The beads were then incubated with 100 mM $FeCl_3$ for 30 minutes with shaking and finally were washed 3 times with 80% acetonitrile/0.1% TFA. Samples were likewise resuspended in 8% acetonitrile/0.15% TFA and incubated with beads for 45 minutes with shaking. The resultant mixture was washed 3 times with 1 mL 80% acetonitrile/0.1% TFA, and eluted using 1:1 acetonitrile:0.7% NH_4OH in water. Eluted phosphopeptides were acidified immediately with 4% formic acid and lyophilized to ~5 μL.

nano-High Performance Liquid Chromatography. For all samples online reverse-phase chromatography was performed using a nanoACQUITY UPLC system (Waters). Peptides were loaded onto a pre-column (75 μm ID, packed with 7 cm C18 particles, Alltech) for 10 min at a flow rate of 1 μL/min. Samples were then eluted over an analytical column (50 μm ID, packed with 15 cm C18 particles, Alltech) using either a 60 or 120 min linear gradient from 2% to 35% acetonitrile with 0.2% formic acid and a flow rate of 300 nL/min.

Target List Construction and *inSeq* Setup. For all experiments, the monoisotopic mass, charge state, and previously determined retention time of target peptides was included for use by the *inSeq* algorithm. In addition, peptides modified on methionines or tyrosines were omitted from all target lists. For peptide elution and isobaric label quantitation *inSeq* experiments, a target list of 4,000 peptides was constructed from a previous nHPLC-MS/MS experiment employing a 90 min nHPLC gradient. For SILAC *inSeq* experiments, peptides identified at 1% FDR in a discovery nHPLC-MS/MS experiment were analyzed to determine the light:heavy partner ratio. A target list of 2,000 peptide pairs (4,000 total peptides) whose ratio deviated from the expected value of 5 by at least 25% was constructed. This subset of peptides included many measurements in which the signal to noise was low, or a partner was missing. For phosphorylation *inSeq* experiments, phosphopeptides identified at 1% FDR in a discovery nHPLC-MS/MS experiment were analyzed by the Phosphinator localization software to assign phosphosite locations. A target list comprising 2,174 phosphopeptides was constructed and used for both ETD only and decision tree (DT) *inSeq* methods.

Target lists were loaded into the instrument's firmware for instant access during acquisition. Peptide lists were stored in an internal database and sorted based on their precursor mass for fast look ups using a binary search algorithm. A parameter file was preloaded into the firmware prior to each experiment to specific scan sequences and instrument parameters needed for the intended experiment.

Mass Spectrometry. All experiments were performed on Thermo LTQ Orbitrap Velos and Q Exactive mass spectrometers. The LTQ Orbitrap Velos used firmware version 2.6.0.1065 SP3 with additional ion trap control language (ITCL) modifications to enable *inSeq* operation. MS scans were performed in the Orbitrap at 30,000 resolution at a max injection time of 500 ms and a target value of 1e6. MS/MS scans were also performed in the Orbitrap at a resolution of 7,500 and with HCD normalize collision energy (NCE) of 27%, for a max fill time of 500 ms. The Q Exactive was operated using version 2.0 Build 142800 with a modified python code base for *inSeq* data acquisition control. Q Exactive MS scans were collected at 70,000 resolution for a max injection time of 120 ms or if the 1e6 AGC target value was reached. MS/MS events were measured at 17,500 resolution at a target value of 1e5, 120 ms max injection time and 26% NCE. Instrument methods for both the LTQ Orbitrap Velos and Q Exactive were overridden during acquisition by the instrument's firmware to provide for dynamic *inSeq* operation. Processing times for *inSeq* were similar on the older Velos-Orbitrap system with the newer Q Exactive; however, a ~100 ms overhead was included because complete collection of the Orbitrap transient signal is necessary before the spectrum can be examined.

Database searching and FDR estimation. MS/MS data was analyzed using the Coon OMSSA Proteomics Software Suite (COMPASS).[42] The Open Mass Spectrometry Search Algorithm (OMSSA; version 2.1.8) was used to search spectra against the International Protein Index(IPI) human database version 3.85.[7] For all experiments, carbamidomethylation of cysteines was included as a fixed modification, while oxidation of methionines was set as a variable modification. For TMT experiments, TMT on the N-terminus and TMT on lysines were included as fixed modifications and TMT on tyrosines was added as a variable modification. For SILAC experiments heavy lysine was added as a variable modification. Precursor mass tolerance was set to ±4.5 Da and monoisotopic mass tolerance was set to ±0.015 Da for fragment ions. Results were filtered to a 1% FDR at both the peptide and protein level with a maximum precursor mass error of 50 ppm. For phosphopeptides, the Phosphinator software was used to localize phosphorylation sites.[30]

Protein and Peptide Quantification. TMT quantification was performed using TagQuant within COMPASS. This program extracts reporter ion intensities and multiplies them by injection times to determine counts. Purity correction was performed as previously described.[43] Tag intensities were normalized to ensure that the total signal from each channel was equal. For evaluation of multiple QuantMode (QM) scans, data was analyzed at the peptide level by only quantifying the first, the sum of first and second, or the sum of the first, second, and third QM scans using TagQuant. Peptides were then combined into protein groups (ProteinHerder) and quantified at the protein level (ProteinTagQuant) within COMPASS. Experimental ratios and p-values (Student's t-test assuming equal variance) were determined using Microsoft Excel. To correct for multiple hypothesis testing, we applied Storey correction using the freely available program QVALUE.[37] SILAC quantification was performed with in-house software that retrieved the peak intensities of both SILAC partners from either a single *inSeq*-triggered SIM scan (monoisotopic peak) or performed an extracted ion chromatogram (30 sec window) of identified precursor. A ratio of partner abundance was only calculated if both SILAC partners had an intensity at least twice that of the noise.

References

[1] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology," *Nature Biotechnology*, vol. 19, no. 3, pp. 242–247, 2001.

[2] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates, A. Bairoch, and J. J. M. Bergeron, "Mass spectrometry in high-throughput proteomics: ready for the big time," *Nature Methods*, vol. 7, no. 9, pp. 681–685, 2010.

[3] C. D. Wenger, G. C. McAlister, Q. W. Xia, and J. J. Coon, "Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass spectrometer," *Molecular & Cellular Proteomics*, vol. 9, no. 5, pp. 754–763, 2010.

[4] A. Michalski, E. Damoc, J. P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning, "Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer," *Molecular & Cellular Proteomics*, vol. 10, no. 9, 2011.

[5] J. K. Eng, A. L. Mccormack, and J. R. Yates, "An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein

database," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[6] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[7] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Y. Yang, W. Y. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–964, 2004.

[8] H. B. Liu, R. G. Sadygov, and J. R. Yates, "A model for random sampling and estimation of relative protein abundance in shotgun proteomics," *Analytical Chemistry*, vol. 76, no. 14, pp. 4193–4201, 2004.

[9] M. R. Hoopmann, G. E. Merrihew, P. D. von Haller, and M. J. MacCoss, "Post analysis data acquisition for the iterative ms/ms sampling of proteomics mixtures," *Journal of Proteome Research*, vol. 8, no. 4, pp. 1870–1875, 2009.

[10] J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nature Methods*, vol. 1, no. 1, pp. 39–45, 2004.

[11] A. Panchaud, A. Scherl, S. A. Shaffer, P. D. von Haller, H. D. Kulasekara, S. I. Miller, and D. R. Goodlett, “Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean,” *Analytical Chemistry*, vol. 81, no. 15, pp. 6481–6488, 2009.

[12] G. C. McAlister, D. H. Phanstiel, J. Brumbaugh, M. S. Westphall, and J. J. Coon, “Higher-energy collision-activated dissociation without a dedicated collision cell,” *Molecular & Cellular Proteomics*, vol. 10, no. 5, 2011.

[13] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, “Higher-energy c-trap dissociation for peptide modification analysis,” *Nature Methods*, vol. 4, no. 9, pp. 709–712, 2007.

[14] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, “Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9528–9533, 2004.

[15] D. P. Little, J. P. Speir, M. W. Senko, P. B. Oconnor, and F. W. Mclafferty, “Infrared multiphoton dissociation of large multiply-charged ions for biomolecule sequencing,” *Analytical Chemistry*, vol. 66, no. 18, pp. 2809–2815, 1994.

[16] C. D. Wenger, M. V. Lee, A. S. Hebert, G. C. McAlister, D. H. Phanstiel, M. S. Westphall, and J. J. Coon, "Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging," *Nature Methods*, vol. 8, no. 11, pp. 933–935, 2011.

[17] J. V. Olsen and M. Mann, "Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 37, pp. 13417–13422, 2004.

[18] D. L. Swaney, G. C. McAlister, and J. J. Coon, "Decision tree-driven tandem mass spectrometry for shotgun proteomics," *Nature Methods*, vol. 5, no. 11, pp. 959–964, 2008.

[19] J. Graumann, R. A. Scheltema, Y. Zhang, J. Cox, and M. Mann, "A framework for intelligent data acquisition and real-time database searching for shotgun proteomics," *Molecular & Cellular Proteomics*, vol. 11, no. 3, 2011.

[20] D. Bailey, G. C. McAlister, C. M. Rose, A. S. Hebert, C. Wenger, M. Lee, M. S. Westphall, and J. J. Coon, "How high mass accuracy measurements will transform targeted proteomics," in *Proceedings of the 59th ASMS Conference on Mass Spectrometry and Allied Topics*, (Denver, Colorado), ASMS, 2011.

[21] W. Yan, J. Luo, M. Robinson, J. Eng, R. Aebersold, and J. Ranish, "Index-ion triggered ms2 ion quantification: A novel proteomics approach for reproducible detection and quantification of targeted proteins in complex mixtures," *Molecular & Cellular Proteomics*, vol. 10, no. 3, 2011.

[22] O. V. Krokhin, R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, and J. A. Wilkins, "An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase hplc - its application to protein peptide mapping by off-line hplc-maldi ms," *Molecular & Cellular Proteomics*, vol. 3, no. 9, pp. 908–919, 2004.

[23] O. V. Krokhin, "Sequence-specific retention calculator. algorithm for peptide retention prediction in ion-pair rp-hplc: Application to 300-and 100-angstrom pore size c18 sorbents," *Analytical Chemistry*, vol. 78, no. 22, pp. 7785–7795, 2006.

[24] R. Kiyonami, A. Schoen, A. Prakash, S. Peterman, V. Zabrouskov, P. Picotti, R. Aebersold, A. Huhmer, and B. Domon, "Increased selectivity, analytical precision, and throughput in targeted proteomics," *Molecular & Cellular Proteomics*, vol. 10, no. 2, 2011.

[25] A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell,

M. Mueller, R. Aebersold, and B. Domon, “An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures,” *Molecular & Cellular Proteomics*, vol. 7, no. 11, pp. 2138–2150, 2008.

[26] A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann, and O. N. Jensen, “Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway,” *Molecular & Cellular Proteomics*, vol. 4, no. 3, pp. 310–327, 2005.

[27] L. Choe, M. D’Ascenzo, N. R. Relkin, D. Pappin, P. Ross, B. Williamson, S. Guertin, P. Pribil, and K. H. Lee, “8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for alzheimer’s disease,” *Proteomics*, vol. 7, no. 20, pp. 3651–3660, 2007.

[28] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther, and M. Mann, “Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast,” *Nature*, vol. 455, no. 7217, pp. 1251–U60, 2008.

[29] K. H. Xiao, J. P. Sun, J. Kim, S. Rajagopal, B. Zhai, J. Villen, W. Haas, J. J. Kovacs, A. K. Shukla, M. R. Hara, M. Hernandez, A. Lachmann, S. Zhao, Y. A. Lin,

Y. S. Cheng, K. Mizuno, A. Ma'ayan, S. P. Gygi, and R. J. Lefkowitz, "Global phosphorylation analysis of beta-arrestin-mediated signaling downstream of a seven transmembrane receptor (7tmr)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 34, pp. 15299–15304, 2010.

[30] D. H. Phanstiel, J. Brumbaugh, C. D. Wenger, S. L. Tian, M. D. Probasco, D. J. Bailey, D. L. Swaney, M. A. Tervo, J. M. Bolin, V. Ruotti, R. Stewart, J. A. Thomson, and J. J. Coon, "Proteomic and phosphoproteomic comparison of human es and ips cells," *Nature Methods*, vol. 8, no. 10, pp. 821–U84, 2011.

[31] M. V. Lee, S. E. Topper, S. L. Hubler, J. Hose, C. D. Wenger, J. J. Coon, and A. P. Gasch, "A dynamic model of proteome changes reveals new roles for transcript alteration in yeast," *Molecular Systems Biology*, vol. 7, 2011.

[32] C. E. Bakalarski, J. E. Elias, J. Villen, W. Haas, S. A. Gerber, P. A. Everley, and S. P. Gygi, "The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses," *Journal of Proteome Research*, vol. 7, no. 11, pp. 4756–4765, 2008.

[33] Y. Zhang, M. Askenazi, J. R. Jiang, C. J. Luckey, J. D. Griffin, and J. A. Marto, "A robust error model for itraq quantification reveals divergent signaling be-

tween oncogenic flt3 mutants in acute myeloid leukemia," *Molecular & Cellular Proteomics*, vol. 9, no. 5, pp. 780–790, 2010.

[34] C. Lengerke, S. Schmitt, T. V. Bowman, I. H. Jang, L. Maouche-Chretien, S. McKinney-Freeman, A. J. Davidson, M. Hammerschmidt, F. Rentzsch, J. B. A. Green, L. I. Zon, and G. Q. Daley, "Bmp and wnt specify hematopoietic fate by activation of the cdx-hox pathway," *Cell Stem Cell*, vol. 2, no. 1, pp. 72–82, 2008.

[35] L. Yang, M. H. Soonpaa, E. D. Adler, T. K. Roepke, S. J. Kattman, M. Kennedy, E. Henckaerts, K. Bonham, G. W. Abbott, R. M. Linden, L. J. Field, and G. M. Keller, "Human cardiovascular progenitor cells develop from a kdr plus embryonic-stem-cell-derived population," *Nature*, vol. 453, no. 7194, pp. 524–U6, 2008.

[36] P. Z. Yu, G. J. Pan, J. Y. Yu, and J. A. Thomson, "Fgf2 sustains nanog and switches the outcome of bmp4-induced human embryonic stem cell differentiation," *Cell Stem Cell*, vol. 8, no. 3, pp. 326–334, 2011.

[37] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[38] S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush, and S. P. Gygi, "A probability-based approach for high-throughput protein phosphorylation analysis and site localization," *Nature Biotechnology*, vol. 24, no. 10, pp. 1285–1292, 2006.

[39] M. J. Schroeder, J. Shabanowitz, J. C. Schwartz, D. F. Hunt, and J. J. Coon, "A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry," *Analytical Chemistry*, vol. 76, no. 13, pp. 3590–3598, 2004.

[40] P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, and R. Aebersold, "High-throughput generation of selected reaction-monitoring assays for proteins and proteomes," *Nature Methods*, vol. 7, no. 1, pp. 43–U5, 2010.

[41] G. K. Chen, D. R. Gulbranson, Z. G. Hou, J. M. Bolin, V. Ruotti, M. D. Probasco, K. Smuga-Otto, S. E. Howden, N. R. Diol, N. E. Propson, R. Wagner, G. O. Lee, J. Antosiewicz-Bourget, J. M. C. Teng, and J. A. Thomson, "Chemically defined conditions for human ipsc derivation and culture," *Nature Methods*, vol. 8, no. 5, pp. 424–U76, 2011.

[42] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "Com-

pass: A suite of pre- and post-search proteomics software tools for omssa," *Proteomics*, vol. 11, no. 6, pp. 1064–1074, 2011.

[43] I. P. Shadforth, T. P. Dunkley, K. S. Lilley, and C. Bessant, "i-tracker: for quantitative proteomics using itraq," *BMC Genomics*, vol. 6, p. 145, 2005.

Chapter 3

INTELLIGENT DATA ACQUISITION BLENDS TARGETED AND DISCOVERY METHODS

Summary

A MS method is described here that can reproducibly identify hundreds of peptides across multiple experiments. The method uses intelligent data acquisition (IDA) to precisely target peptides while simultaneously identifying thousands of other, non-targeted peptides in a single nano-LC-MS/MS experiment. We introduce an online peptide elution order alignment (EOA) algorithm that targets peptides based on their relative elution order, eliminating the need for retention time-based scheduling. We have applied this method to target 500 mouse peptides across six technical replicate nano-LC-MS/MS experiments and were able to identify 440 of these in all six, compared to only 201 peptides using data-dependent acquisition (DDA). A total of 3,757 other peptides were also identified within the same experiment, illustrating that this hybrid method does not eliminate the novel discovery advantages of DDA. The method was also tested on a set of mice in biological quadruplicate and increased the number of identified target peptides in all four mice by over 80% (826

vs. 459) compared with the standard DDA method. We envision real-time data analysis as a powerful tool to improve the quality and reproducibility of proteomic datasets.

Introduction

Large-scale proteomic studies make use of a variety of tools and techniques to achieve depth and wide coverage of proteomes. The most popular method for sequencing proteomes is shotgun sequencing where peptides are digested from extracted proteins, separated with chromatography (HPLC), and then mass analyzed using mass spectrometry (MS).[1,2] Since complex proteomes can encompass thousands of proteins, leading to millions of peptides, deciding how to allocate the limited mass spectrometer bandwidth is key to successful analysis.[3] By far the most successful method for this time management is data dependent acquisition (DDA), where intact peptide precursors are first mass analyzed (MS), specific *m/z* features are then selected to undergo fragmentation, and finally the fragment ions are mass analyzed again (MS/MS). This process is repeated throughout the LC separation, resulting in a large collection of MS and MS/MS spectra. Peptides are eventually identified from the fragmentation spectra and then assembled together back into protein groups.[4–8] This approach has produced outstanding results in the past decade, but, due to variety of reasons (e.g., large protein dynamic range, speed of MS instrumentation, separation efficiency, etc.) undersampling of proteomes is very common. In other words, not every peptide is identified in every LC-MS/MS experiment. Incomplete datasets limit the questions researchers can answer; especially when biological replication is used to increase statistical power, many measurements become worthless if they cannot be measured reproducibly.[9] As proteomics seeks to answer global biological questions, reproducible peptide identification between datasets is mandated.[10–12]

Many studies have outlined the problem of poor peptide reproducibility.[13–17] Aebersold succinctly summarized that irreproducibility is a multifaceted issue, depending on user experience, equipment, and data analysis, among others.[18] He outlines that there are two main approaches in tackling irreproducibility. First, exhaustively identify every peptide in a sample—an approach that is becoming more feasible as technology improves.[19–21] The more common approach, as many other researchers have embarked on, is to focus on a smaller subset of peptides and to thoroughly identify and quantify those using targeted methods.[22] Methods such as selected reaction monitoring (SRM) are powerful and reproducible, but are low throughput, targeting a few hundred peptides at most.[23–27] To improve identification reproducibility and throughput, targeted methods almost exclusively rely on retention time-based scheduling, segmenting the MS duty cycle among the target peptides. In SRM methods, a series of MS/MS transitions for each targeted peptide is automatically collected at the appropriate retention time (RT), removing the dependence on MS detection. This requires precise knowledge of the peptide retention time for the LC-MS system and is low throughput as only one set of transitions are monitored at a given point in time. Recent work on intelligent SRM (iSRM) increases throughput by monitoring only a subset of transitions for each target, switching to normal SRM when these transitions are detected.[28] We sought to expand upon the idea of intelligent real-time switching of methods by combining the enhanced reproducibility of targeted scheduled methods with the novel discovery advantages of DDA in a single hybrid method. Our goals were three-fold: first, to develop a method that increases the throughput of targeting; second, to replace retention-time based scheduling and its laborious method development with a more robust and straightforward peptide elution ordering; and last, to maintain the discovery aspect of DDA sampling while simultaneously targeting a subset of peptides.

In the last decade, a few computational approaches have been aimed at solving the problem of poor reproducibility. The concept of accurate mass tags (AMT) was first introduced by Smith et. al. as a means to identify peptides in multiple runs based on accurate mass and retention time.[29] This concept was further expanded with PepMiner and PEPPeR, tools for clustering features among multiple datasets.[30,31] Most notably, Prakash et. al. introduce the concept of aligning multiple MS datasets based on peptide relative elution order into signal maps.[32] To date, these and other computational methods[33–38] have been performed post-acquisition, attempting to improve already collected data. We seek to improve the reproducibility at the source by improving the algorithms the MS uses to select precursors to fragment. We and others have proposed using real-time data analysis and dynamic MS control as a means for improving the quality of acquired spectra.[39–41] Here we present our findings on combining accurate mass, elution orders, and real-time data analysis to improve the sampling reproducibility of the MS.

Results and Discussion

Irreproducible Peptide Identification. In data dependent acquisition (DDA) peptide precursors are selected for fragmentation based on intensity in a MS survey scan. This straightforward approach has proven to be a simple and powerful technique. However, it is pestered with inconsistent sampling, and therefore, irregular peptide identification between experiments. The DDA method is inherently stochastic in nature, depending heavily on the consistency of the input data (MS) to deliver reproducible peptide identification (MS/MS). Even the slightest change in the chromatography or ionization efficiencies will have repercussions on the collection of the whole dataset (e.g., the butterfly effect). To characterize the extent these minor changes have on the reproducibility of peptide identifications, six replicate injections of a tryptic digest of yeast whole cell lysate were analyzed using DDA on the same nano-LC-MS/MS system over a span of ten days. On average, each experiment identified 13,289 ±340 unique peptide sequences (I/L ambiguity removed) at a 1% FDR, indicating a highly consistent separation and nearly identical instrument performance. Of the 23,919 unique peptides identified in total, only 5,404 (22.6%) of those peptide were identified in all six experiments (Figure 3.1). A significant portion were only identified once (7,474 31.2%) while the remaining peptides were divided between two and five experiments. This clearly demonstrates the irreproducibility of DDA sampling on the same peptide solution. The reproducibility of identified protein groups fares better; 1,708 of 3,054 (56%) protein groups were identified in every experiment. The higher overlap percentage is because many different peptides can make up one protein group, minimizing the importance of identifying the same peptides in all experiments. However, post translational modification (PTM) analysis requires identification of the same sites to compare between experiments, demanding the need for high peptide overlap. PTM analysis

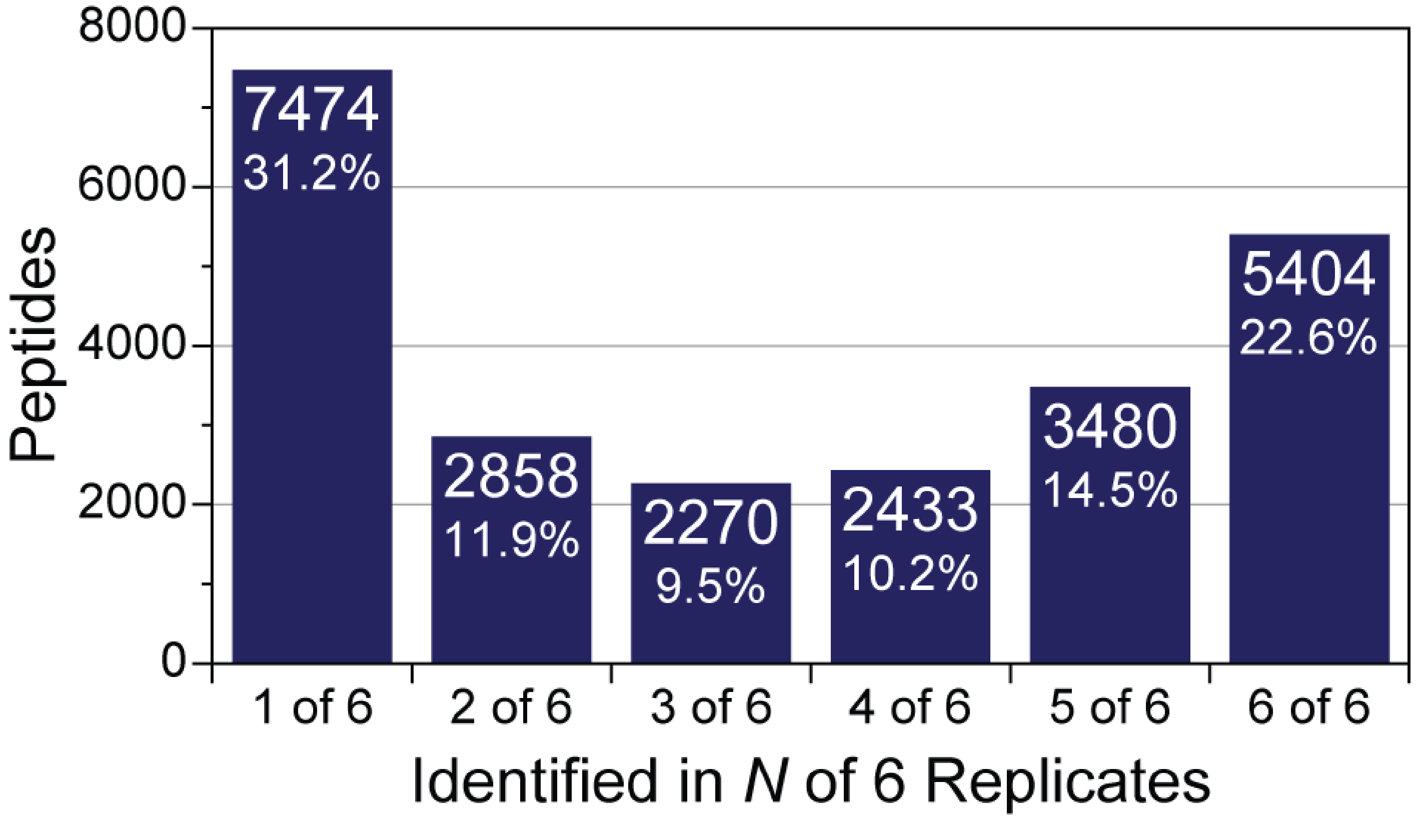


Figure 3.1: Overlap of peptide identification among the analysis of six technical replicates. Six nano-LC-MS/MS experiments produced 23,919 unique peptide identifications in total, but only one fifth of the identifications were observed in all six replicates. A large percentage (31.2%) of the peptides were only detected in one of the six experiments.

and quantitation is becoming more prominent in the literature, thus making this a growing problem in the field. Two reasons can be attributed to the poor reproducibility of stochastic DDA sampling. First, precursors having low signal-to-noise (S/N) are affected first by changes in chromatography and ionization. For example, a precursor with a maximal S/N of 4 may have been sampled and identified in one experiment, but in the next experiment the S/N may have dropped below the detection threshold and excluded from being sampled. This is evident when 8,883 MS features from peptides identified in one or all of the six experiments were examined for their maximal S/N (Figure 3.2). For peptides identified once 2,707 (30.5%) had a maximal S/N $\leqslant$ 4 while only 814 (9.2%) precursors identified in every experiment had similar maximum S/N. The other reason for inconsistent peptide identification is increased MS spectral complexity, specifically its effect on charge-state assignment. In proteomic MS/MS workflows, precursors are often only selected when they exhibit a well-defined charge state—usually where $z > 1$, as singly charged precursors fragment poorly and usually do not lead to positive identifications. Increases in spectral complexity hinder the charge-state determination algorithms, especially for low S/N precursors. This results in skipping precursors even if its signal-to-noise is above the sampling threshold.

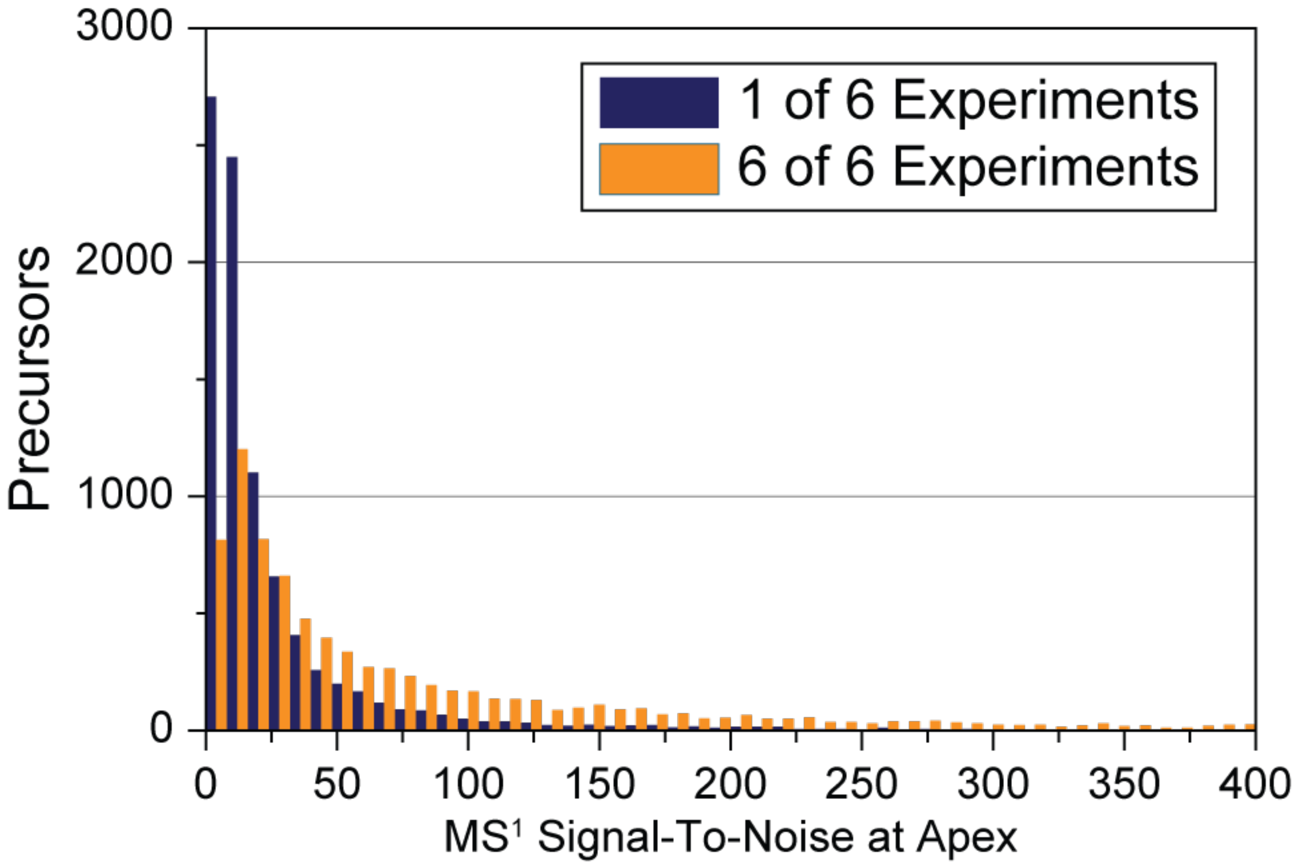


Figure 3.2: Distribution of signal-to-noise ratios of reproducible and irrepoducible peptide identifications. Peptides that were identified in 1 of 6 or 6 of 6 DDA top-15 experiments were analyzed for their maximal MS signal-to-noise (S/N). A larger percentage of those peptides seen only once appear at lower signal-to-noise values, indicating that MS signal intensity for reproducible identifications.

Retention Time Based Targeting. When good peptide identification reproducibility is needed, retention time (RT) based targeting, i.e., scheduling, has been the method of choice. Here, peptides of interest are assigned an expected elution time and MS/MS are triggered, regardless of MS detection, during the appropriate time range. This avoids the two issues with DDA sampling described above and enables much higher reproducibility. However, such methods are laborious to construct and maintain; identical LC and MS parameters must be kept between experiments to minimize any variances in retention times of the peptides.

To assess the degree of variance in peptide retention times that occur in normal nano-LC-MS/MS experiments, two of the yeast DDA experiments described above, performed ten days apart, were compared. The first experiment (July 22nd, D0) produced 13,529 unique peptides and the second experiment (July 31st, D9) identified 13,433 yeast peptides. Together, 7,589 peptides were in common and the apex of their retention time in each experiment is plotted in Figure 3.3A.

The relationship between retention times of matched peptides is highly linear ($R^2 = 0.9989$) but has a non-unity slope and non-zero intercept ($m = 1.033$; $b = -0.647$). While the slope is very close to 1, even the slightest deviation (0.033), compounded over time, leads to large RT differences late in the separation (e.g., ~1.6 min shift at 70 min). On the whole, the average RT deviation was nearly a minute ($\mu = -0.805$

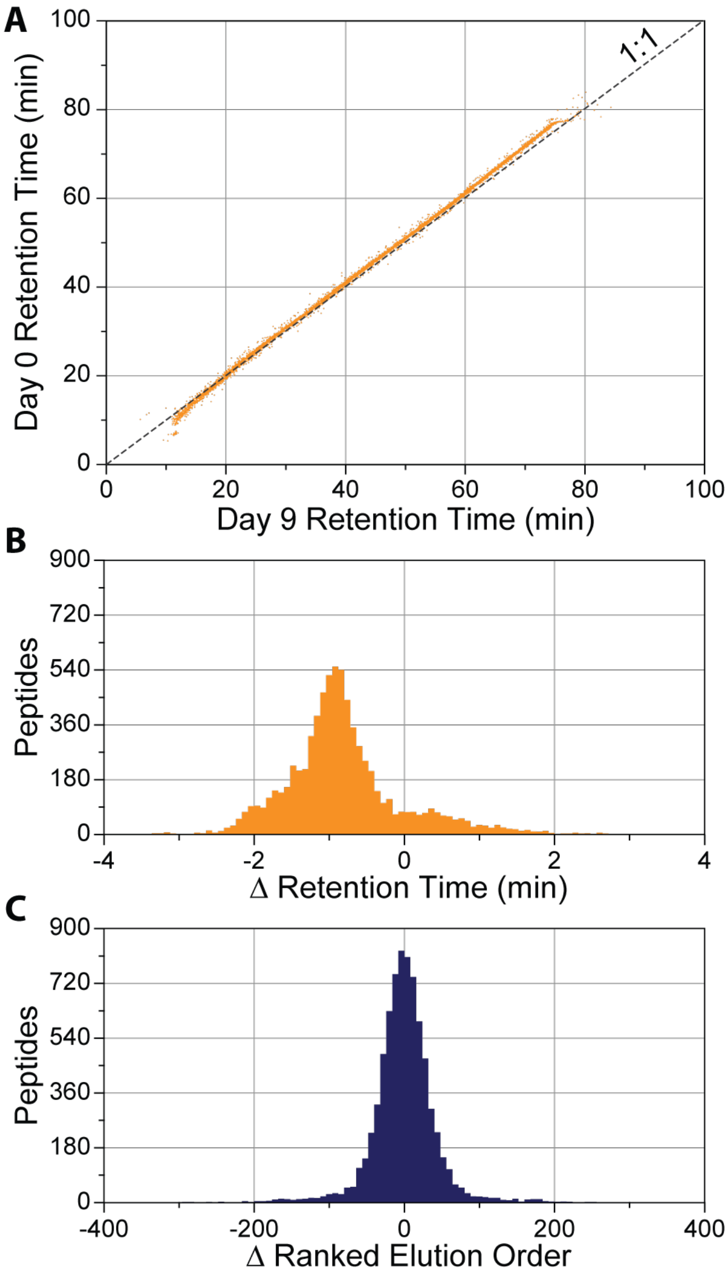


Figure 3.3: Retention time deviation between matched LC-MS/MS experiments. To assess the deviation in retention times for matched samples two identical nano-LC-MS/MS experiments were run ten days apart on the same LC-MS system. (A) The relationship between apex retention times of the 7,589 unique peptides common between experiments display a high degree of linearity (R^2 = 0.9989) but a skewed slope and non-zero intercept (m = 1.033; b = -0.647). (B) The average deviation from unity was nearly a minute off (μ = -0.805 min), with a broad distribution over 2 minutes wide. (C) Peptides ranked by their relative elution order exhibit a normal distribution around zero (μ =-1.097).

min) with a broad distribution over a two minute range (Figure 3.3B). Typically, the assigned peptide elution times must be corrected to encompass this shift.

We hypothesize that—due to the degree of linearity in peptide retention times,we could avoid these corrections by scheduling peptides based on their relative elution order (EO), opposed to their absolute retention time. Under similar LC conditions (i.e., same particles, temperature, column length, phase, etc.) peptides elute in the same relative order regardless of separation duration or slope. For example, if peptide 'A' elutes before peptide 'B' in a 30 minute LC gradient, the same ordering is preserved with a 60 minute LC gradient, even if the absolute retention times vary greatly. When many peptides' elution orders are taken into account (e.g., 1000s of peptides) they provide a simple way to correct for elution variation dynamically. This is evident when we took the 7,589 peptides and rank ordered them based on their apex retention times for both the D0 and D9 experiments and plotted the difference between matched peptides (Figure 3.3C). Here the values are normally distributed around zero (μ =-1.097) with a full width at half maximum (FWHM) of only ~100.

Elution order can be useful even under extreme differences in chromatographic conditions as well. To simulate dynamic chromatographic conditions, we separated yeast peptides under two different LC gradient profiles. The resulting peptide iden-

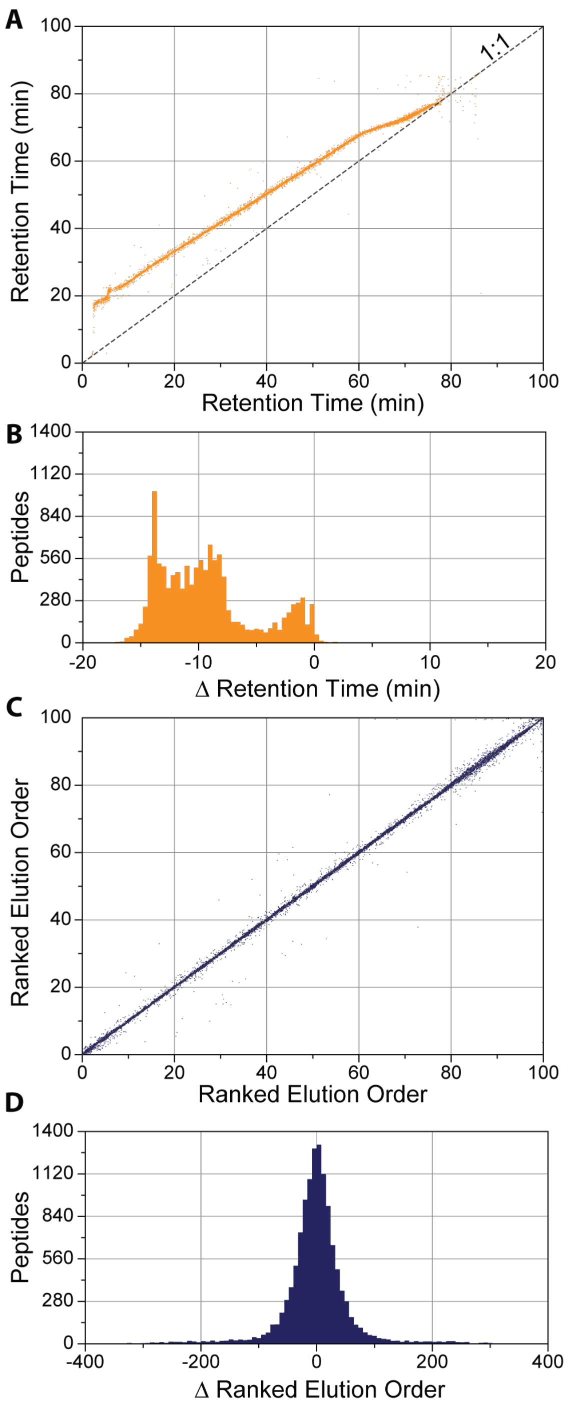


Figure 3.4: Two technical replicates of a yeast DDA top-15 method using two different LC gradients. (A) Retention times for matched peptides between the two gradients is linear but not 1:1. (B) An average deviation of 10 min exists between the two experiments. (C) Rank elution orderings of matched peptides between the two gradients show a highly linear relationship that is exactly 1:1. (D) The average deviation in elution order for matched peptides is symmetric around 0.

tifications were again matched between the runs and the retention time difference was plotted (Figure 3.4A). These data show an average deviation of ten minutes between the two gradients (Figure 3.4B), but when ranked by their elution orders, the two experiments show a linear slope of 1 with a normal distribution of ranked elution orders around zero (Figure 3.4C&D).

Real-time Elution Ordering Alignment. We reasoned that using elution order could improve the irreproducible sampling of DDA—similarly to scheduled methods, but on a larger scale and more robustly. The question shifts from ”What retention time is it?” as scheduled methods ask, to ”What is the current elution order?” By knowing which peptides are currently eluting from the LC, combined with the *a priori* knowledge of their elution order, we predict with high fidelity which peptides are going to subsequently elute.

Prior knowledge is needed of the sample to adequately calculate the elution orders of the peptides in the sample. With time-based scheduled methods, many cursory experiments are performed to optimize the retention times of the targeted peptides. To reduce variances in retention times, it's vital that these initial experiments are conducted exactly the same as the targeted experiments. In stark contrast, elution orders can be determined using a variety of methods. First, much work has been devoted to determining peptide hydrophobicities from theoretical calculations of the amino acid sequence.[42–45] A simple list of peptides, ordered by their hydrophobicities, can produce a highly linear elution ordering. Second, previously collected data of the sample can produce an accurate elution ordering as long as the LC conditions are similar enough. This enables the combination of multiple datasets to produce a single elution order vs. *m/z* map (elution order map, EOM), regardless of their individual separation durations. This is accomplished by rank ordering all the peptide identifications in a given run and normalizing their orderings between 0 and 100 (where 100 represents the last eluting peptide). These normalized values are then matched between experiments and aligned using a simple algorithm to produce the final EOM as shown in Figure 3.5A. Lastly, the most robust method for determining peptide elution orders is to perform a discovery experiment right before the targeted experiment. Regardless of how elution order is determined, the final EOM is uploaded onto the instrument and is accessed throughout the course of the subsequent analyses.

Prior to targeted analysis, a list of peptide targets, along with their relative elution orders are also uploaded to the instrument (Figure 3.6B). Each target is assigned an elution order range depending on how long it was identified in the discovery experiments (see Figure 3.6C for zoom in). During the targeted analysis, instead of relying on absolute retention time to trigger targeted MS/MS, determining the

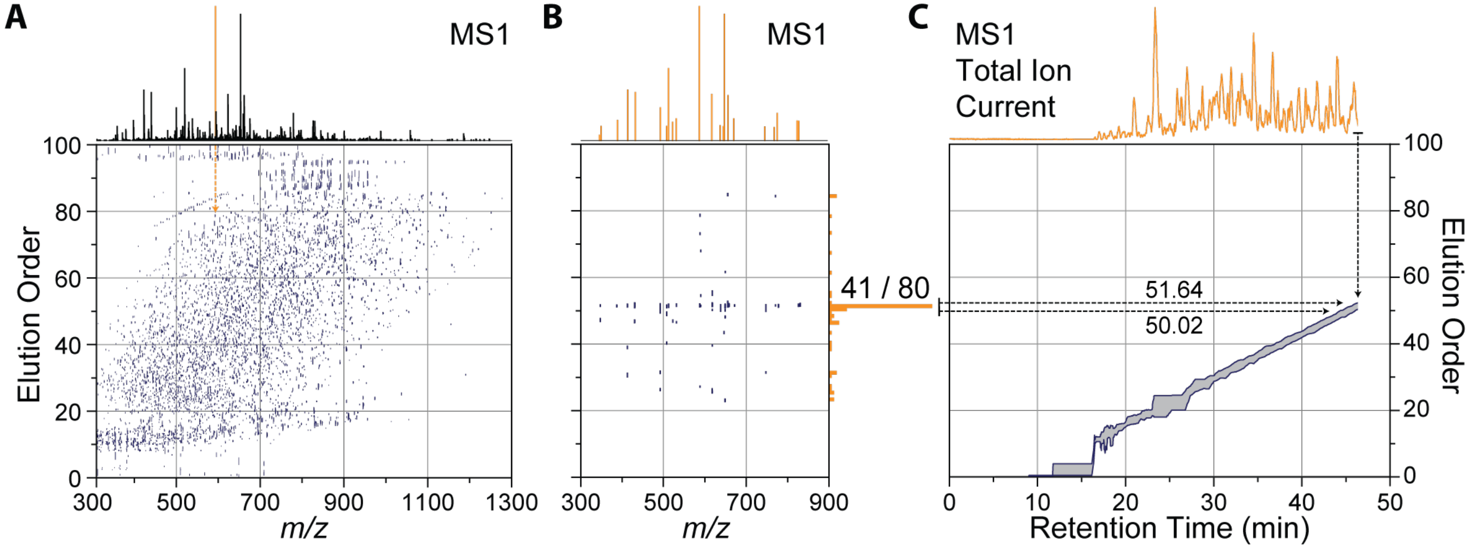


Figure 3.5: Real-time elution order alignment algorithm. After 46.3 minutes into a LC-MS/MS experiment, an MS scan is performed (A) and *m/z* features are matched to a 2D ion map stored on the instrument. (B) 21 of the peaks match 80 features in the ion map at a 10 ppm tolerance. Of these, over half (41 of 80) were mapped to one elution order bin (51 elution order). (C) A rolling elution order range is continually updated throughout the LC-MS/MS experiment.

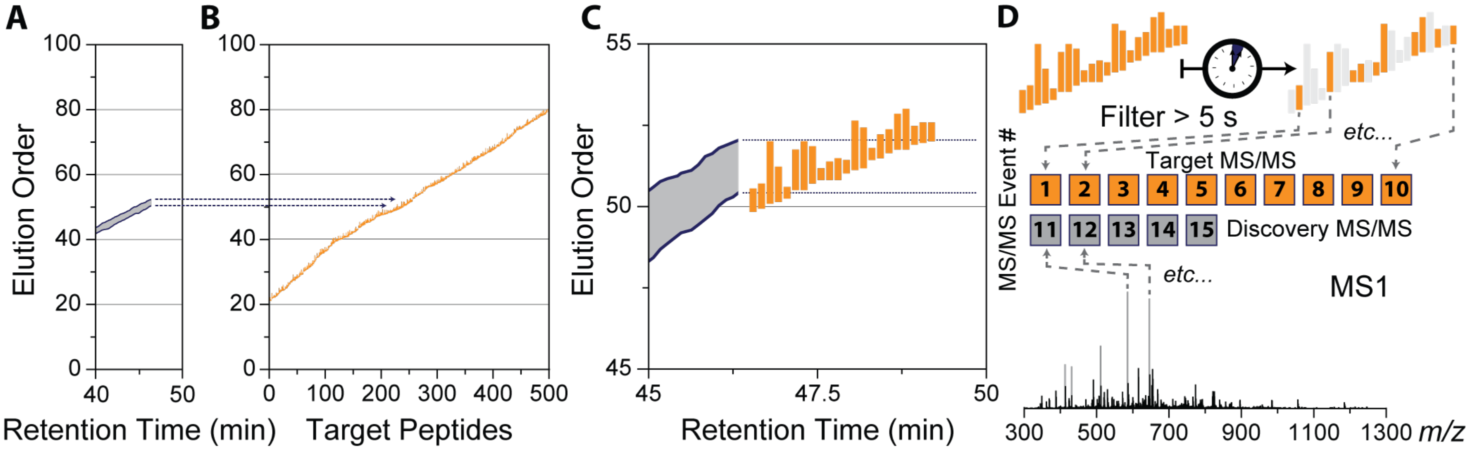


Figure 3.6: Selection of targets within elution order range. Following determination of the current elution order range (A), targets sharing a similar elution order are selected (C). Peptide targets within the elution order range are sorted based on when they were last sampled for MS/MS, leaving targets that have been waiting the longest. Those peptides are immediately sampled, regardless of MS detection (D). Remaining MS/MS events are automatically filled with DDA chosen *m/z* features.

current elution order becomes the main goal of the method. We have designed an online peptide elution order alignment (EOA) algorithm that takes a single MS spectrum and computes the current elution order therefrom. In brief, following MS acquisition, the EOA algorithm takes the most intense *m/z* feature and extracts all the elution order values from the uploaded EOM at a narrow *m/z* tolerance (e.g., 10 ppm) (Figure 3.5A). Each *m/z* feature is matched in a similar fashion and the resulting EO values are stored in a separate array (Figure 3B). In this example MS, 21 *m/z* features matched a total of 80 EO values. When analyzed, 41 of these values are contained within a single 1 EO-wide bin. This indicates with high confidence that the current elution order is near this maximum. To determine the elution order precisely, the algorithm then calculates the 95% confidence interval around the max EO bin and stores the minimum (50.02) and maximum (51.64) elution order. This process is repeated for each MS and over time the calculated elution order range constructs a rolling-average as shown in Figure 3C. The EOA algorithm is expedient, taking on average 26 ms per MS to execute and does not induce a statistically significant change in the total number of MS/MS scans performed (Figure 3.7).

Once the current elution order range is determined, peptides sharing a similar elution order are selected for MS/MS analysis. Briefly, the current elution order

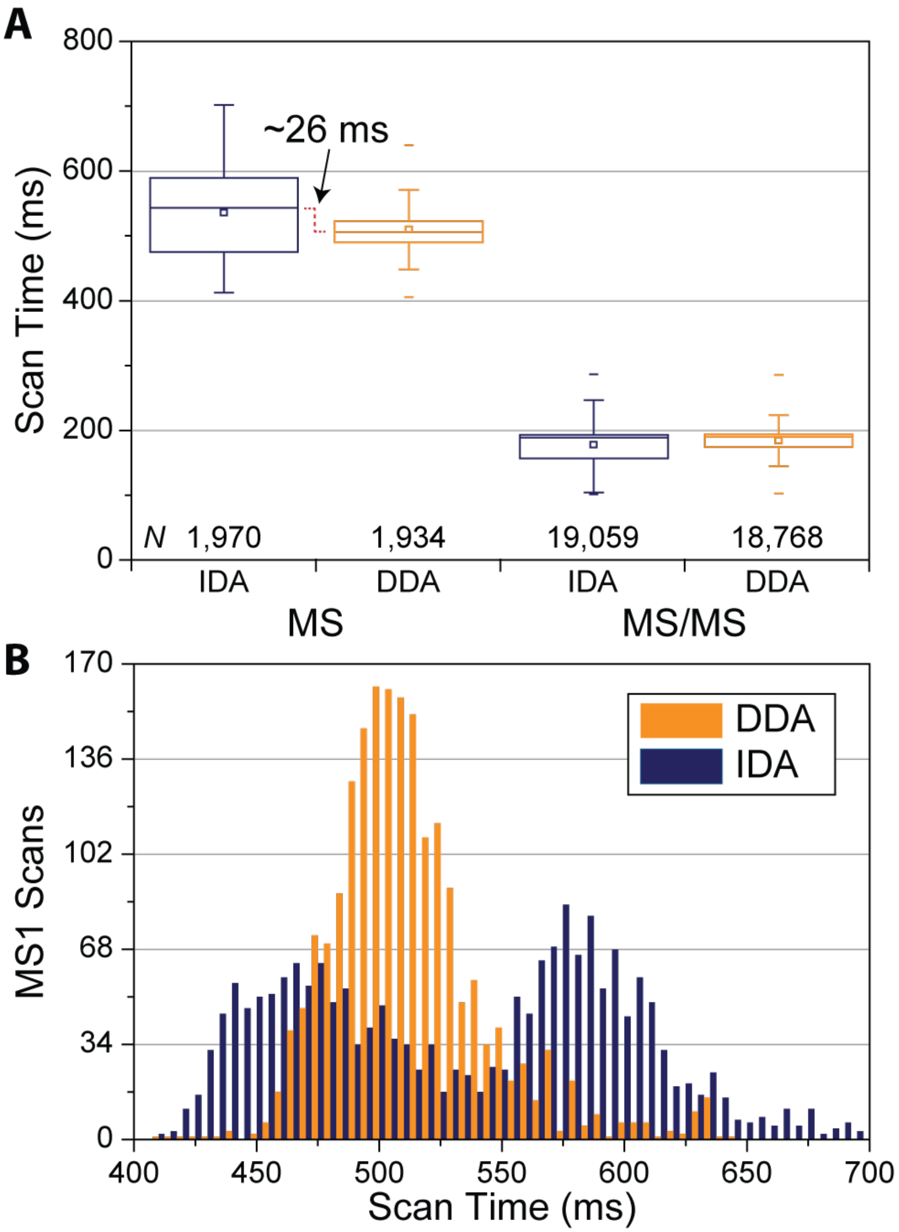


Figure 3.7: Duty cycle of elution order alignment algorithm. (A) The elution order alignment (EOA) algorithm is expedient and induces only a slight increase in the MS duty cycle compared to normal DDA method (~26 ms). (B) The distributions of scan times for IDA is bimodal because the EOA algorithm can be triggered every other MS, because the current elution order changes only slightly between consecutive MS scans.

range is intersected with the target peptides already uploaded on the instrument (Figure 3.6B) and overlapping peptides are stored as potential targets (Figure 3.6C). These peptides have a high probability of eluting next since they share very similar EO values with the current overall EO value. Since there can be many potential targets at any given time, they are filtered based on how long since they were last sampled, this is to prevent oversampling of any one target. Peptides that have been waiting the longest (i.e., > 5s) are automatically triggered for MS/MS analysis regardless of MS detection. Unfilled MS/MS events are then populated using normal DDA top-N approaches, excluding any *m/z* previously selected to be targeted (Figure 3.6D). This data collection scheme enables repetitive, consistent targeting of multiple peptides over their elution, while allowing DDA scans to facilitate discovery. The EOA algorithm is compatible with other quantitative strategies such as parallel reaction monitoring (PRM) where peptide targets are repeatedly sampled (MS/MS) over their elution, and the resulting fragment ions are extracted to provide quantitative information (Figure 3.8).[46,47]

Improving Peptide Identification in Multiple Experiments. We reasoned that the EOA algorithm would improve the reproducibility of peptide identification across multiple runs. Additionally, we increased the proteomic complexity by using a mammalian system (mouse) to determine how complexity affects the algorithm.

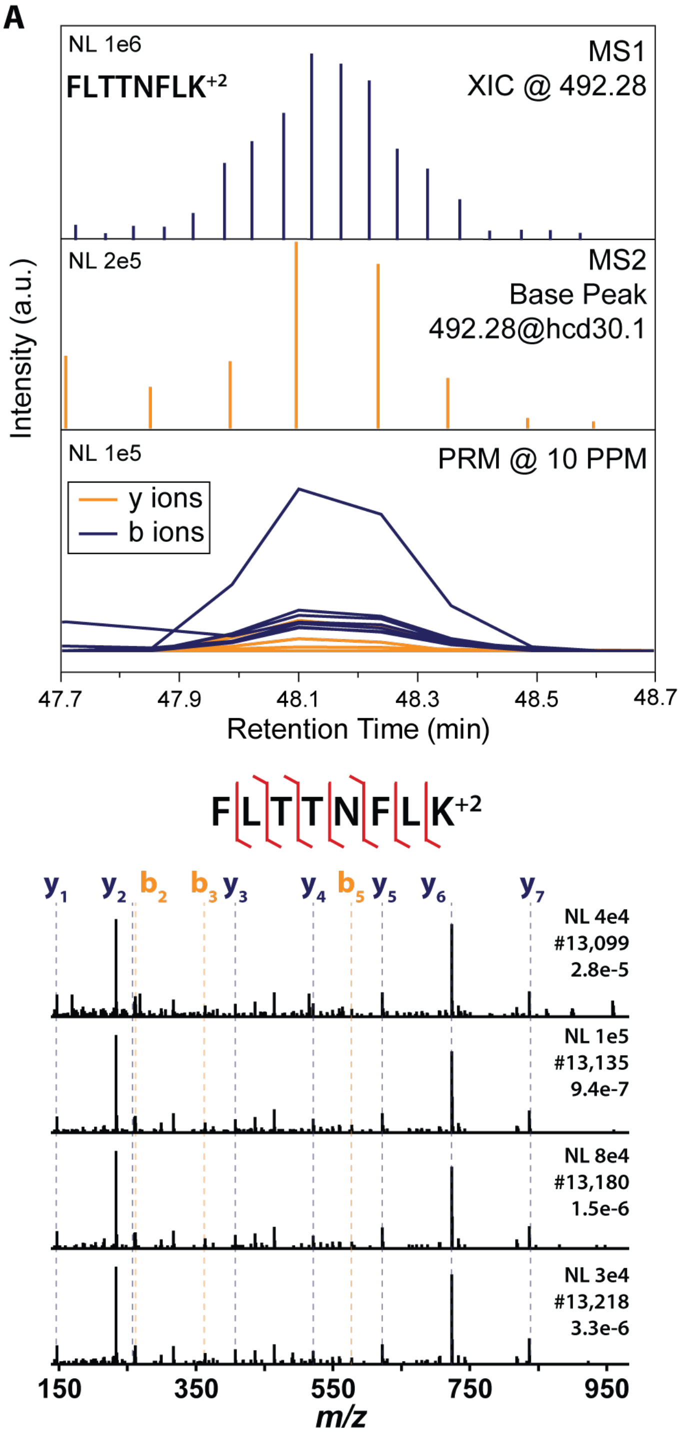


Figure 3.8: Parallel reaction monitoring (PRM) scan sequences obtainable using the IDA method. (A) The peptide FLTTNFLK was MS/MS sampled approximately every 6 seconds over its elution profile. The b- and y-ions intensities were tracked over time to provide quantitative results.

To test its effectiveness, 500 mouse peptides—identified in only three of six previous discovery experiments, were randomly selected to serve as peptide targets. Each peptide's elution order was calculated from the discovery experiments they were identified in, combined into a single EOM, and then uploaded to the instrument (Figure 3.6B). The same vial of peptides was then repeatedly injected and successively analyzed using DDA, an inclusion list (INC), and intelligent data acquisition (IDA) in hexplicate. On average, only 251 (50%) peptides were identified using DDA (Figure 3.9A, 1% FDR, error bars represent one σ). This is consistent with the discovery data where the selected target peptides originated from three of six experiments (50%). The accurate mass inclusion list modestly increases identifications to 280 (56%) but the biggest improvement is realized with IDA, where 440 of 500 targets (88%) were identified on average. Since the IDA method enables simultaneous un-targeted MS/MS sampling, comparisons of the total number of peptide identifications between the three acquisition methods can be made (Figure 3.9B). Each method produced nearly the same number of peptide spectral matches (PSMs). A difference appears at the unique PSMs level (i.e., peptides) where both DDA and INC produced similar number of identifications (~5,800 peptides) but dropped to ~3,700 using IDA. We attributed this decline primarily to the redundant sampling of target peptides with the IDA method compared to the other methods.

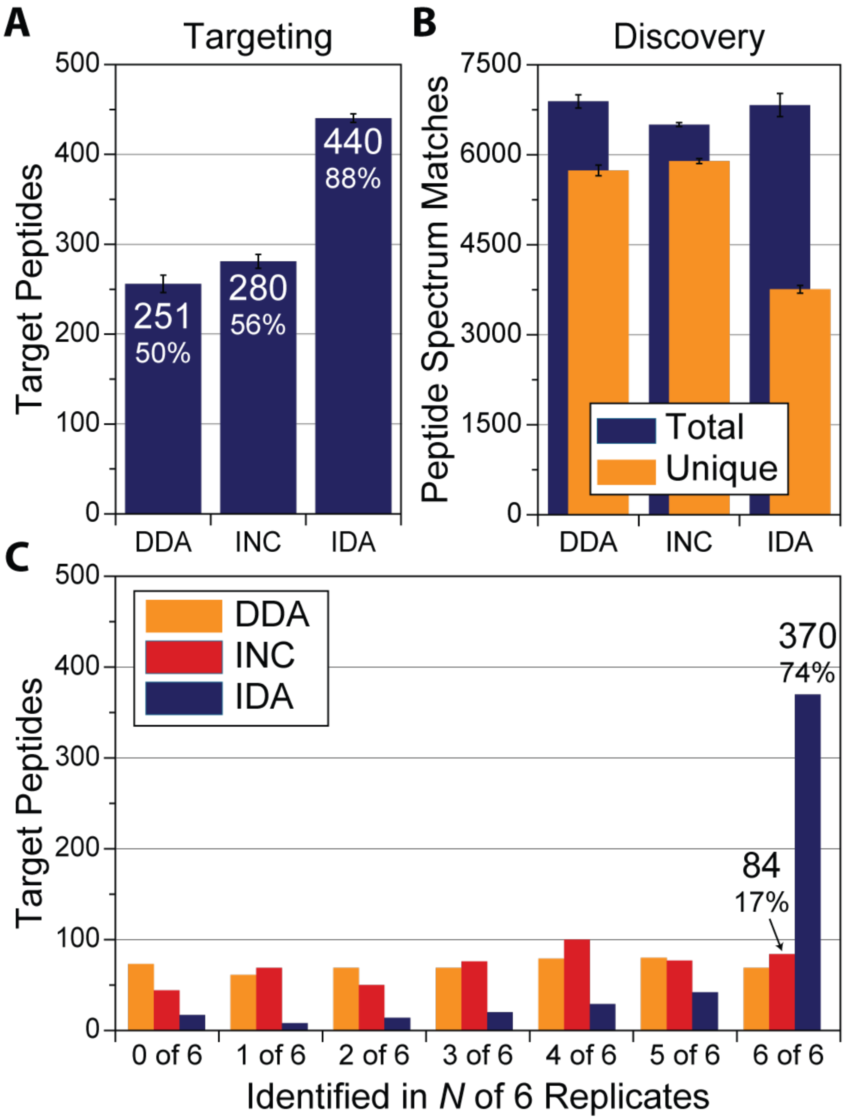


Figure 3.9: Reproducibility improvements using intelligent data acquisition. A subset of 500 mouse peptides were targeted with DDA, an accurate mass inclusion list (INC), and our intelligent data acquisition (IDA) method in hexplicate. (A) IDA identified the most target peptides of the three methods (error bars represent 1 σ). (B) Discovery identifications by three methods show only a slight decline in the total number of peptides identified using IDA. (C) 74% of the targets were observed in all six technical replicates when IDA was used compared to less than 20% for the inclusion list or data dependent acquisition.

IDA identified each target 4.3 times on average, compared to 0.59 and 0.63 for DDA and INC respectively, a ~7:1 ratio. This is in agreement with the ratio of dynamic exclusion times between methods; IDA uses 5 seconds for each target, compared to the longer dynamic exclusion time (35 s, 1:7) used in the DDA and INC methods. The oversampling of target peptides in IDA increases the likelihood of identification. We feel that it is an acceptable tradeoff between maximizing reproducibility for a subset of peptides and a slight decline in total identified peptides. The increased reproducibility is demonstrated in Figure 5C; the IDA method identified 370 (74%) of the same peptides in all six experiments. The same cannot be said for DDA or INC, which only managed to identify 69 and 84 peptides in all six experiments, respectively. This represents an increase of over 340% in the number of peptide targets that were seen in all replicates.

Improved Reproducibility in Biological Systems. All data described above have consisted of technical replicates of the same sample, injected with the same HPLC, and analyzed using the same MS. These technical replicates are ideal to develop acquisitions methods on, primarily because the same peptides should exist in each injection, which removes sample variability from obfuscating the results. However, biological replication in proteomic studies is becoming more prevalent, due to the increase in statistical power it affords. To test whether intelligent data acquisition improves reproducibility in biological systems, four male C57BL/B6 mice were sacrificed at ten weeks, eight organs were harvested, and peptides from a tryptic digestion of each organ was labeled with a TMT 8-plex tag (Figure 3.10A&B). The tagged peptides from each mouse were mixed together and separated over a 165 minute gradient and sampled using a DDA top-15 method to generate a list of peptide targets. An average of 8,683 ±313 peptide sequences were identified in each mouse for a total of 13,502 unique sequences. Of these, only 3,969 (29.4%) peptides were identified in every mouse (Figure 3.10C). A subset of 1,500 peptides were selected from the peptides detected in either two or three of four mice and sorted based on their assigned elution orders (Figure 3.10D). Here, peptide targets were chosen to be evenly distributed in the elution order dimension to limit the number of coeluting peptides at a given point. In subsequent targeting experiments, each mouse sample was analyzed twice, once using DDA and the other IDA, for a total of eight experiments. When the DDA targeting experiments were analyzed, an average of 810 (54%) target peptides were identified (Figure 3.11A, 1% FDR, error bars represent one σ). Using IDA, this number increases to 1,072 (71.5%).

In total, over half of the targeted peptides (826, 55.1%) were identified in all four mice when using IDA compared to only 30.6% (459) using DDA (Figure 3.11B). The IDA method represents a nearly 80% improvement over DDA in the number of

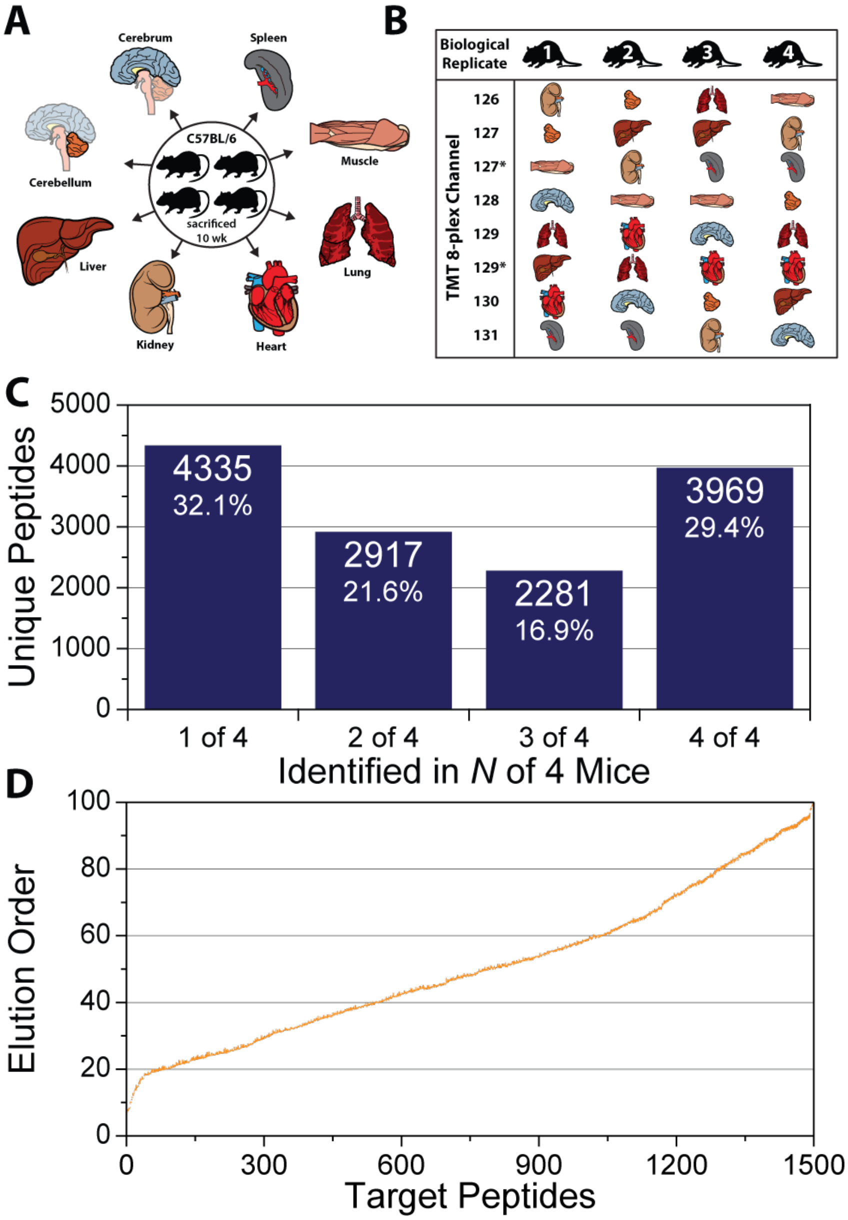


Figure 3.10: Peptide targets of biologicial replicates of mice. (A) Four C57BL/6 mice were sacrificed at 10 weeks of age and eight organs were harvested from each mouse. (B) Peptides resulting from a tryptic digestion of lysates from each organism were labeled with TMT 8-plex tags in a randomized order. (C) The breakdown of peptides identified in the four mice using DDA top-15 method. A small percentage (29.4%) were only seen in all four mice. (D) A subset of 1,500 peptide targets were selected from peptides only detected in 2 or 3 of all 4 mice.

peptide targets it identifies in all mice. This increase in reproducible identification improves the quantitative results as well. When each tissue is compared to liver, the number of peptides that could be statistically quantified (p-value < 0.05) is on average 227 greater with IDA compared to DDA (Figure 3.11C). For example, when the quantitative data for muscle is compared to liver (Figure 3.11D), IDA produced 826 significantly different peptides while only 531 were significant for DDA, a 56% increase. This can be directly attributed to increased reproducibility in identification across biological samples.

Conclusion

The ability to identify the same peptides in multiple experiments reproducibly is increasingly important in proteomic analysis as increased statistical power is demanded. Historically—the most common acquisition method, data-dependent acquisition (DDA) has been used to sample large portions of proteomes, but lacks adequate peptide identification reproducibility. In this manuscript, we expand upon our previous intelligent data acquisition (IDA) work and introduce the concept of using elution order as a way to schedule and target peptides. Here we have described an online elution order alignment (EOA) algorithm that automatically adjusts to different chromatographic conditions to deliver consistent scheduling

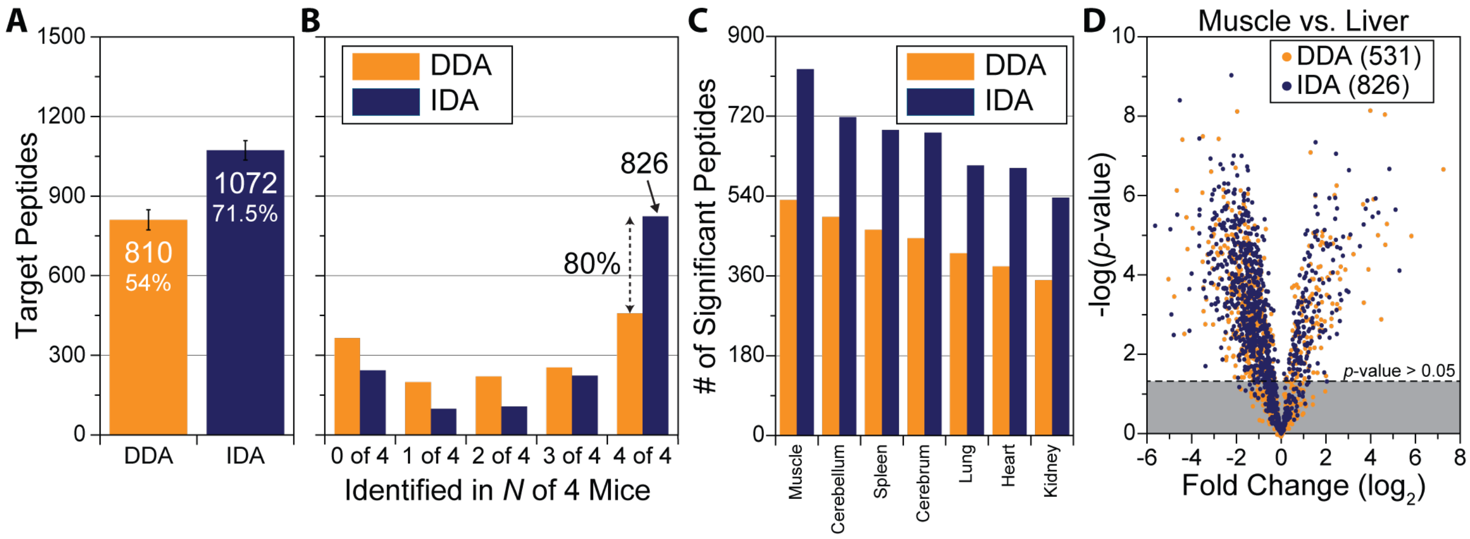


Figure 3.11: Identification and quantification improvements in biological replication. (A) In four subsequent nano-LC-MS/MS experiments, only 810 of 1,500 mouse peptide targets were identified with DDA. The identifications improve to 1,072 when IDA is used (error bars represent 1 Ïƒ). (B) In total, 826 target peptides were identified in all four mouse when IDA was used to target. This number falls to only 459 peptides when DDA is used. (C) The number of statistically significant differences (p-value < 0.05) quantified when each tissue is compared to liver is greater with IDA than DDA. (D) When comparing target peptides identified in the muscle vs. the liver, IDA quantified 826 statistically significant peptides compared to only 531 when DDA was used, a 56% increase.

and robust reproducibility. The method is capable of targeting large numbers of peptides (>500) in a single run with minimal upfront preparation and effort. Using this method, we have shown great improvements in peptide identification overlap among multiple experiments compared to DDA (88% compared to 50% identification overlap in six experiments). The EOA algorithm is capable of improving reproducibility even for highly variable samples. In four mice, our method was able to identify 806 target peptides compared to only 459 using normal DDA sampling.

We believe that such technologies can now be applied to traditional SRM methods that use triple quadrupole mass spectrometers. Here, periodic full MS scans could be performed and analyzed to calculate the current elution order and adjust the timing of the SRM transitions. One challenge would be the decreased specificity in determining elution order from low resolution scans. However, using a more adaptable metric for scheduling (elution ordering vs. retention time) could potentially increase the portability and robustness of SRM methods while reducing development time.

A novel aspect of our method is the combination of discovery and targeted analysis in a single method. The MS intelligently switches between targeted and discovery modes depending on what is currently eluting, without any human intervention. In one experiment, over 3,700 unique mouse peptides were discovered using DDA while simultaneously targeting 500 peptides. Such hybrid MS methods enable both a focused and holistic view on the same sample, something that is welcomed when sample-limited.

Until comprehensive proteomic coverage is routinely obtained, targeted methods will be heavily used and developed. We have explored increasing the intelligence of MS methods as a means to improve the throughput and power of peptide targeting, without sacrificing the novel discovery aspect of DDA sampling. Future work includes improvements to the determination of elution orders, increasing the success rate of target identification, and maximizing the throughput to target larger portions of the proteome without laborious upfront work.

Experimental Methods

Yeast Culture. *Saccharomyces cerevisiae* strain BY4741 was grown in yeast extract peptone dextrose media (YPD) (1% yeast extract, 2% peptone, 2% dextrose). A starter culture was added to 2 L of media and was propagated for ~12 generations (20 hours) to a total OD_{600} of ~2. The cells were pelleted with centrifugation at 5,000 rpm for 5 min, supernatant decanted, and re-suspended in chilled NanoPure water. Washing with water was repeated twice and the final pelleting was performed at 5,000 rpm for 10 min. The pellet was resuspended in lysis buffer composed of

50 mM Tris pH8, 8 M urea, 75 mM sodium chloride, 100 mM sodium butyrate, protease and phosphatase inhibitor tablet (Roche). Cell lysing was performed with glass bead milling in a stainless steel container (Retsch). A 2.5 mL aliquot of resuspended yeast were shaken with 2 mL of acid-washed glass beads at 30 Hz for 4 min, followed with 1 min rest, for eight cycles.

Mouse Handling and Tissue Isolation. Four male C57BL/B6 mice were bred from in-house colonies and housed in an environmentally controlled facility with free access to water and standard rodent chow (Purina #5008). Mice were kept in accordance to the University of Wisconsin-Madison Research Animals Resource Center and NIH guidelines for care and use of laboratory animals. At 10 weeks of age, mice were sacrificed by decapitation after a four hour fast. Eight tissues were dissected from the mice (cerebellum, cerebrum, kidney, heart, liver, lung, extensor digitorum longus, and spleen), flash frozen in liquid nitrogen and stored at -80°C. Tissues were homogenized in 1 mL lysis buffer/100 mg tissue (8 M Urea, 50 mM Tris, 100 mM NaCl, 1 mM $CaCl_2$, 100 mM sodium butyrate, 5 μM MS-275, 0.2 μM SAHA, Roche protease and phosphatase inhibitor tablets).

Sample Preparation. Protein was quantified by BCA (Pierce) and reduced with 5 mM dithiothreitol and incubated for 45 minutes at 55°C. Alkylation was performed with 15 mM iodoacetamide for 30 minutes in the dark and quenched with 5 mM dithiothreitol. Urea concentration was diluted to 1.5 M with 50 mM Tris pH 8.0. Proteolytic digestion was performed by addition of Trypsin (Promega), 1:50 enzyme to protein ratio, and incubated at ambient temperature overnight. For quantitative studies, the resulting peptides were labeled with TMT 8-plex (Pierce) isobaric tag, and mixed.[48,49] All samples were desalted using C-18 solid phase extraction (SPE) columns (Waters, Milford, MA) prior to nano-LC-MS/MS analysis.

Nano LC-MS/MS analysis. Peptides were separated with online reverse-phase chromatography using a nanoACQUITY UPLC system (Waters, Milford, MA). Peptides were first loaded onto a precolumn (75 μm ID, 5 cm Magic C18 particles, Bruker, Michrom) for 10 min at 1 μl/min flow rates. Peptides were then separated on a 30 cm analytical column (75 μm ID, 5 cm Magic C18 particles) for either 100 or 160 min over a linear gradient from 8% to 35% acetonitrile at 300 nl/min. Mass analysis was performed on an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, San Jose, CA) using 60,000 resolving power (RP) MS scans.[50] Peptides selected for MS/MS analysis used a 2 Th isolation width, fragmented with HCD (NCE = 35), and analyzed in the Orbitrap at 15,000 RP or 30,000 RP for quantitative experiments. Unless otherwise noted, data-dependent analysis was performed selecting the top 15 most intense *m/z* features (charge state >1) for MS/MS analysis. Dynamic exclusion settings were enabled for 35 s at ±10 ppm mass window, 1 occurrence with a maximum of 500 exclusions at any given point in time. Automatic Gain Control (AGC) was enabled and MS targets were set to 1×10^6 and MS/MS targets were set to 5×10^4. Accurate mass inclusion list experiments would prioritize MS/MS sampling from a list of targets at ±10 ppm mass tolerances. Remaining MS/MS events were filled with normal top-N DDA approaches. Intelligent data acquisition control was implemented using the ion trap control language (ITCL, Thermo Fisher Scientific). Briefly, following MS analysis, the spectra was analyzed using algorithms written in ITCL to select targets for MS/MS analysis (described herein). Any remaining MS/MS slots would be filled by the unmodified DDA firmware code.

Data Analysis. Thermo .raw files were processed using the Coon OMSSA Proteomic Analysis Software Suite (COMPASS) and in-house software.[51] Briefly, raw files were converted to the dta file format (DTA Generator) and were searched using the Open Mass Spectrometry Search Algorithm (OMSSA, v 2.1.9).[52] Yeast data was searched against a target-decoy database of yeast ORFs (www.yeastgenome.com, February 3, 2011) and mouse data from UniProt canonical database.[53] Peptides were generated from a tryptic digestion with up to three missed cleavages, carbamidomethylation of cysteines as fixed modifications, and oxidation of methionines as variable modifications. For quantitative experiments, a fixed modification of 8-plex TMT tag was added to lysines and peptide n-terminus, with a variable modification of 8-plex TMT tag on tyrosines. Precursor mass tolerance was 100 ppm using the multiisotope function (-tem 4 -ti 4) and product ions were searched at 0.015 Da tolerances. Peptide spectral matches (PSM) were validated using FDR Optimizer based on q-values at a 1% false discovery rate (FDR). Protein groups were constructed from peptide identifications according to the law of parsimony and filtered to a 1% FDR (Protein Hoarder). For quantitative datasets, peptides were quantified with TagQuant (v1.4) using the generated TMT 8-plex reporter ions, corrected for isotopic impurities, and normalized to total protein abundance. Peptide Elution orders determination algorithms were performed by custom software developed in C# with the Microsoft .NET Framework version 4.5.

References

[1] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology," *Nat Biotechnol*, vol. 19, no. 3, pp. 242–7, 2001.

[2] D. A. Wolters, M. P. Washburn, and J. R. Yates, "An automated multidimensional protein identification technology for shotgun proteomics," *Anal Chem*,

vol. 73, no. 23, pp. 5683–90, 2001.

[3] A. Michalski, J. Cox, and M. Mann, "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc-ms/ms," *J Proteome Res*, vol. 10, no. 4, pp. 1785–93, 2011.

[4] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J Am Soc Mass Spectrom*, vol. 5, no. 11, pp. 976–89, 1994.

[5] R. G. Sadygov, D. Cociorva, and J. R. Yates, "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book," *Nat Methods*, vol. 1, no. 3, pp. 195–202, 2004.

[6] J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nat Methods*, vol. 1, no. 1, pp. 39–45, 2004.

[7] M. R. Hoopmann, G. E. Merrihew, P. D. von Haller, and M. J. MacCoss, "Post analysis data acquisition for the iterative ms/ms sampling of proteomics mixtures," *J Proteome Res*, vol. 8, no. 4, pp. 1870–5, 2009.

[8] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data:

the protein inference problem," *Mol Cell Proteomics*, vol. 4, no. 10, pp. 1419–40, 2005.

[9] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal Bioanal Chem*, vol. 389, no. 4, pp. 1017–31, 2007.

[10] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science*, vol. 292, no. 5518, pp. 929–34, 2001.

[11] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[12] M. P. Molloy, E. E. Brzezinski, J. Hang, M. T. McDowell, and R. A. VanBogelen, "Overcoming technical variation and biological variation in quantitative proteomics," *Proteomics*, vol. 3, no. 10, pp. 1912–9, 2003.

[13] H. Liu, R. G. Sadygov, and J. R. Yates, "A model for random sampling and estimation of relative protein abundance in shotgun proteomics," *Anal Chem*, vol. 76, no. 14, pp. 4193–201, 2004.

[14] A. Wolf-Yadlin, S. Hautaniemi, D. A. Lauffenburger, and F. M. White, “Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks,” *Proc Natl Acad Sci U S A*, vol. 104, no. 14, pp. 5860–5, 2007.

[15] D. L. Tabb, L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A. J. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. A. Carr, K. R. Clauser, J. D. Jaffe, K. A. Kowalski, T. A. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman, “Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry,” *J Proteome Res*, vol. 9, no. 2, pp. 761–76, 2010.

[16] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates, A. Bairoch, and J. J. Bergeron, “Mass spectrometry in high-throughput proteomics: ready for the big time,” *Nat Methods*, vol. 7, no. 9, pp. 681–5, 2010.

[17] F. Pachl, B. Ruprecht, S. Lemeer, and B. Kuster, “Characterization of a high field orbitrap mass spectrometer for proteome analysis,” *Proteomics*, vol. 13, no. 17, pp. 2552–62, 2013.

[18] R. Aebersold, “A stress test for mass spectrometry-based proteomics,” *Nat Methods*, vol. 6, no. 6, pp. 411–2, 2009.

[19] S. S. Thakur, T. Geiger, B. Chatterjee, P. Bandilla, F. Frohlich, J. Cox, and M. Mann, “Deep and highly sensitive proteome coverage by lc-ms/ms without prefractionation,” *Mol Cell Proteomics*, vol. 10, no. 8, p. M110 003699, 2011.

[20] N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm, and M. Mann, “System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra hplc runs on a bench top orbitrap,” *Mol Cell Proteomics*, vol. 11, no. 3, p. M111 013722, 2012.

[21] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, “The one hour yeast proteome,” *Mol Cell Proteomics*, 2013.

[22] M. M. Savitski, F. Fischer, T. Mathieson, G. Sweetman, M. Lang, and M. Bantscheff, “Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays,” *J Am Soc Mass Spectrom*, vol. 21, no. 10, pp. 1668–79, 2010.

[23] V. Lange, P. Picotti, B. Domon, and R. Aebersold, “Selected reaction monitoring for quantitative proteomics: a tutorial,” *Mol Syst Biol*, vol. 4, p. 222, 2008.

[24] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold, "Full dynamic range proteome analysis of s. cerevisiae by targeted proteomics," *Cell*, vol. 138, no. 4, pp. 795–806, 2009.

[25] P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, and R. Aebersold, "High-throughput generation of selected reaction-monitoring assays for proteins and proteomes," *Nat Methods*, vol. 7, no. 1, pp. 43–6, 2010.

[26] P. Picotti and R. Aebersold, "Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions," *Nat Methods*, vol. 9, no. 6, pp. 555–66, 2012.

[27] E. Sabido, Y. Wu, L. Bautista, T. Porstmann, C. Y. Chang, O. Vitek, M. Stoffel, and R. Aebersold, "Targeted proteomics reveals strain-specific changes in the mouse insulin and central metabolic pathways after a sustained high-fat diet," *Mol Syst Biol*, vol. 9, p. 681, 2013.

[28] R. Kiyonami, A. Schoen, A. Prakash, S. Peterman, V. Zabrouskov, P. Picotti, R. Aebersold, A. Huhmer, and B. Domon, "Increased selectivity, analytical precision, and throughput in targeted proteomics," *Mol Cell Proteomics*, vol. 10, no. 2, p. M110 002931, 2011.

[29] R. D. Smith, G. A. Anderson, M. S. Lipton, L. Pasa-Tolic, Y. Shen, T. P. Conrads, T. D. Veenstra, and H. R. Udseth, “An accurate mass tag strategy for quantitative and high-throughput proteome measurements,” *Proteomics*, vol. 2, no. 5, pp. 513–23, 2002.

[30] I. Beer, E. Barnea, T. Ziv, and A. Admon, “Improving large-scale proteomics by clustering of mass spectrometry data,” *Proteomics*, vol. 4, no. 4, pp. 950–60, 2004.

[31] J. D. Jaffe, D. R. Mani, K. C. Leptos, G. M. Church, M. A. Gillette, and S. A. Carr, “Pepper, a platform for experimental proteomic pattern recognition,” *Mol Cell Proteomics*, vol. 5, no. 10, pp. 1927–41, 2006.

[32] A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, and B. Schwikowski, “Signal maps for mass spectrometry-based comparative proteomics,” *Mol Cell Proteomics*, vol. 5, no. 3, pp. 423–32, 2006.

[33] D. Radulovic, S. Jelveh, S. Ryu, T. G. Hamilton, E. Foss, Y. Mao, and A. Emili, “Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry,” *Mol Cell Proteomics*, vol. 3, no. 10, pp. 984–97, 2004.

[34] J. Listgarten and A. Emili, “Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry,” *Mol Cell Proteomics*, vol. 4, no. 4, pp. 419–34, 2005.

[35] Y. Shen, E. F. Strittmatter, R. Zhang, T. O. Metz, R. J. Moore, F. Li, H. R. Udseth, R. D. Smith, K. K. Unger, D. Kumar, and D. Lubda, “Making broad proteome protein measurements in 1-5 min using high-speed rplc separations and high-accuracy mass measurements,” *Anal Chem*, vol. 77, no. 23, pp. 7763–73, 2005.

[36] H. Zhang, E. C. Yi, X. J. Li, P. Mallick, K. S. Kelly-Spratt, C. D. Masselon, n. Camp, D. G., R. D. Smith, C. J. Kemp, and R. Aebersold, “High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry,” *Mol Cell Proteomics*, vol. 4, no. 2, pp. 144–55, 2005.

[37] H. Lin, L. He, and B. Ma, “A combinatorial approach to the peptide feature matching problem for label-free quantification,” *Bioinformatics*, vol. 29, no. 14, pp. 1768–75, 2013.

[38] N. W. Bateman, S. P. Goulding, N. Shulman, A. K. Gadok, K. K. Szumlinski, M. J. Maccoss, and C. C. Wu, “Maximizing peptide identification events in pro-

teomic workflows utilizing data-dependent acquisition," *Mol Cell Proteomics*, 2013.

[39] D. J. Bailey, C. M. Rose, G. C. McAlister, J. Brumbaugh, P. Yu, C. D. Wenger, M. S. Westphall, J. A. Thomson, and J. J. Coon, "Instant spectral assignment for advanced decision tree-driven mass spectrometry," *Proc Natl Acad Sci U S A*, vol. 109, no. 22, pp. 8411–6, 2012.

[40] J. Graumann, R. A. Scheltema, Y. Zhang, J. Cox, and M. Mann, "A framework for intelligent data acquisition and real-time database searching for shotgun proteomics," *Mol Cell Proteomics*, vol. 11, no. 3, p. M111 013185, 2012.

[41] J. T. Webber, M. Askenazi, S. B. Ficarro, M. A. Iglehart, and J. A. Marto, "Library dependent lc-ms/ms acquisition via mzapi/live," *Proteomics*, vol. 13, no. 9, pp. 1412–6, 2013.

[42] O. V. Krokhin, R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, and J. A. Wilkins, "An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase hplc: its application to protein peptide mapping by off-line hplc-maldi ms," *Mol Cell Proteomics*, vol. 3, no. 9, pp. 908–19, 2004.

[43] O. V. Krokhin, “Sequence-specific retention calculator. algorithm for peptide retention prediction in ion-pair rp-hplc: application to 300- and 100-a pore size c18 sorbents,” *Anal Chem*, vol. 78, no. 22, pp. 7785–95, 2006.

[44] K. Petritis, L. J. Kangas, B. Yan, M. E. Monroe, E. F. Strittmatter, W. J. Qian, J. N. Adkins, R. J. Moore, Y. Xu, M. S. Lipton, n. Camp, D. G., and R. D. Smith, “Improved peptide elution time prediction for reversed-phase liquid chromatography-ms by incorporating peptide sequence information,” *Anal Chem*, vol. 78, no. 14, pp. 5026–39, 2006.

[45] V. Spicer, M. Grigoryan, A. Gotfrid, K. G. Standing, and O. V. Krokhin, “Predicting retention time shifts associated with variation of the gradient slope in peptide rp-hplc,” *Anal Chem*, vol. 82, no. 23, pp. 9678–85, 2010.

[46] A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall, and J. J. Coon, “Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics,” *Mol Cell Proteomics*, vol. 11, no. 11, pp. 1475–88, 2012.

[47] S. Gallien, E. Duriez, C. Crone, M. Kellmann, T. Moehring, and B. Domon, “Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer,” *Mol Cell Proteomics*, vol. 11, no. 12, pp. 1709–23, 2012.

[48] G. C. McAlister, E. L. Huttlin, W. Haas, L. Ting, M. P. Jedrychowski, J. C. Rogers, K. Kuhn, I. Pike, R. A. Grothe, J. D. Blethrow, and S. P. Gygi, “Increasing the multiplexing capacity of tmts using reporter ion isotopologues with isobaric masses,” *Anal Chem*, vol. 84, no. 17, pp. 7469–78, 2012.

[49] T. Werner, I. Becher, G. Sweetman, C. Doce, M. M. Savitski, and M. Bantscheff, “High-resolution enabled tmt 8-plexing,” *Anal Chem*, vol. 84, no. 16, pp. 7188–94, 2012.

[50] A. Michalski, E. Damoc, O. Lange, E. Denisov, D. Nolting, M. Muller, R. Viner, J. Schwartz, P. Remes, M. Belford, J. J. Dunyach, J. Cox, S. Horning, M. Mann, and A. Makarov, “Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down lc ms/ms and versatile peptide fragmentation modes,” *Mol Cell Proteomics*, vol. 11, no. 3, p. O111 013698, 2012.

[51] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, “Compass: a suite of pre- and post-search proteomics software tools for omssa,” *Proteomics*, vol. 11, no. 6, pp. 1064–74, 2011.

[52] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, “Open mass spectrometry search algorithm,” *J Proteome Res*, vol. 3, no. 5, pp. 958–64, 2004.

[53] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” *Nat Methods*, vol. 4, no. 3, pp. 207–14, 2007.

Chapter 4

SOFTWARE FRAMEWORKS FOR PROTEOMIC DATA ANALYSIS

Summary

Proteomic research can be divided into three major parts: 1) sample generation and preparation, 2) mass spectra collection, and 3) data interpretation and analysis. While improvements in both sample preparation and instrumentation have greatly propelled the field forward, data analysis software has developed at a slower rate. This may be a result of competing standards of data storage and access, the shear complexity of large-scale data, or the simple fact that a majority of scientists are not programmers. Whatever the case may be, automatic data analysis is needed to help answer large and meaningful biological problems. The existence of software is not the final goal; the tools must be simple to use yet powerful, flexible yet robust, accessible yet timely in order to gain traction and be impactful to the field. To meet these demands, the following chapter describes the development of two open-source software packages used in proteomic data analysis. The first package is the Coon OMSSA Proteomic Analysis Software Suite (COMPASS), a graphical interface program used to analyze proteomic data from initial spectral processing all the way to protein quantitation. It is geared for the end user to process their data in a

straightforward, but flexible manner. The second package is devoted to developing new software tools in a timely fashion and is called C# Mass Spectrometry Library (CSMSL). This programming toolbox offers a wide range of proteomic and mass spectrometry tools and methods for developing new software analysis tools quickly. It is powerful to handle the most complex data, but approachable that even novice programmers can use it with minimal training. These two software tools are still in their infancy, but are constantly being updated and maintained to meet the needs of the ever-changing proteomic landscape.

Introduction

In many scientific disciplines, as the complexity of the problems grow, so to do the informatic resources to keep pace. Proteomics and mass spectrometry are no exception. As researchers aim to answer larger biological problems on grander scales, proteomic data analysis needs to keep up. The mass spectrometers used to collect the data are becoming faster and more sensitive (i.e., more data) with each passing year. Acquiring 20 MS/MS spectra per second is now possible. These mass spectrometers are also becoming more robust and powerful, enabling them to be continually run for several consecutive days with very little down time. In the end, hundred of thousands of spectra are collect per instrument in an average day, totaling millions over a week. Keeping up with this volume of data requires sophisticated software and data management tools. These tools must be powerful to handle the complexity of the data, flexible to changes and updates in how data is analyzed, and simple to use that non-programmers can effectively use them.

To meet these requirements, I have developed a range of software tools to analyze mass spectrometry data. First, the Coon Research group has previously published a suite of software tools called Coon OMSSA (Open Mass Spectrometry Search Algorithm) Proteomic Analysis Software Suite (COMPASS). The suite encompasses all the basic tools for analyzing mass spectrometry-based proteomic data. It handles the spectral cleaning and conversion, and supports MS/MS searching (*via* OMSSA)[1], false discovery analysis, protein grouping, various types of quantitation, post-translational modification localization, and protein quantitation, among others. To handle the massive number of spectra collected by a host of users per day, I have utilized the High Throughput Condor (HTCondor) system—developed here at the University of Wisconsin-Madison, to greatly speed up the database searching program by simultaneously using hundreds of computers across campus. Since its initial publication, COMPASS has been heavily upgraded and expanded to meet the changing needs of the group. This chapter summarizes the different programs of COMPASS and the changes and updates made to them since the publication of the software.

The second software tool I have developed is an open source programming library specifically designed for proteomic data analysis called C# Mass Spectrometry Library (CSMSL). This library speeds up the development of new software, providing many common functions and concepts needed for data analysis (e.g., peptides, chemical formulas, spectral searching, etc.). This removes the burden from the programmer to reinvent the wheel each time data needs to be analyzed. Each of its features have been carefully designed and tested to provide a powerful, flexible, and robust set of tools for programmers to use. We kept the design of the library as simple as possible, enabling even novice programmers to quickly analyze their data in unique fashions with little training. Given the complexity of proteomics and mass spectrometry, this allows the user to focus more on the scientific data than the management and construction of complex software. In this chapter, the concept and design of CSMSL will be outlined and a few coding examples are provided to show the simplicity of the library. CSMSL is an ongoing work, the version at the time of this publication (v0.2.1) is far from its final form.

COMPASS: Coon OMSSA Proteomic Analysis Software Suite

The COMPASS program is a complete, standalone data analysis platform for proteomic mass spectrometry. It is based around the Open Mass Spectrometry Search Algorithm (OMSSA) as the primary MS/MS search engine, but has been partially adapted to handle inputs from other proteomic search engines (e.g., SEQUEST and Proteome Discoverer).[2] It is written for the Windows operating system using their .NET Framework (v3.5 and above) in the C# programming language. The complete source code is available at `https://github.com/dbaileychess/Compass` and the version is v1.2.12 at the time of this publication. The application contains a graphical user interface (GUI), making it very intuitive and easy to use. COMPASS contains several other GUI programs, each corresponding to a separate step in the analysis. Users process data files through each individual program, which typically writes the result of the analysis to different files and folders on the computer. Other programs then process those result files to add additional analyses and outputs. This design enables customized analysis workflows to handle the various types of analyses commonly used (Figure 4.1).

COMPASS was first published in 2011, but has been substantially upgraded to improve the user experience, fix bugs, introduce new features, and speed up its execution. The program was also heavily refactored (code reorganization) to

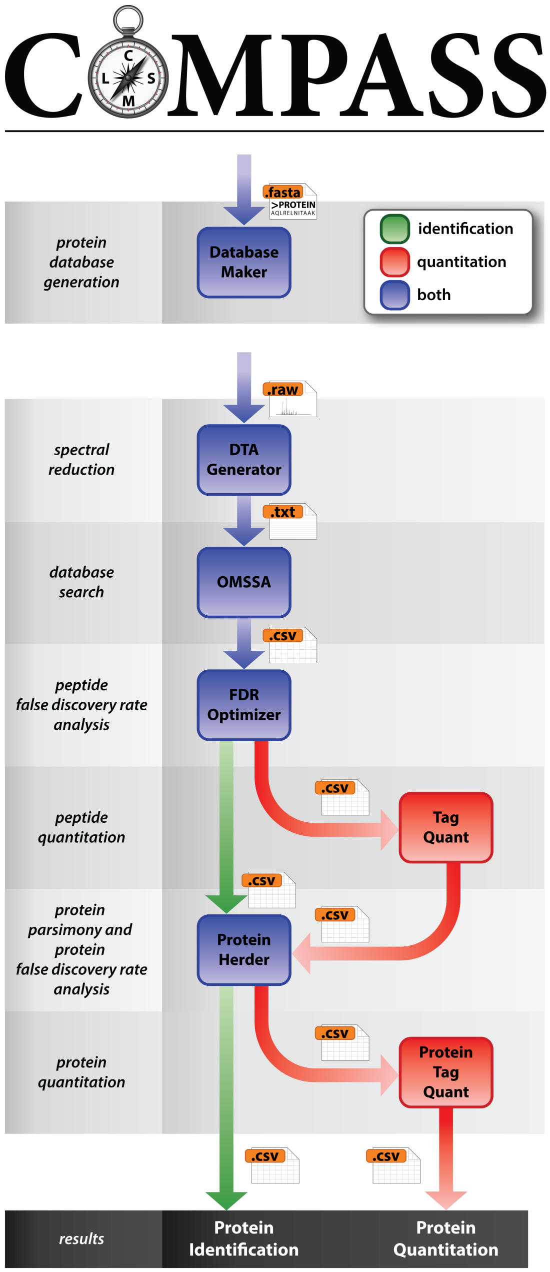
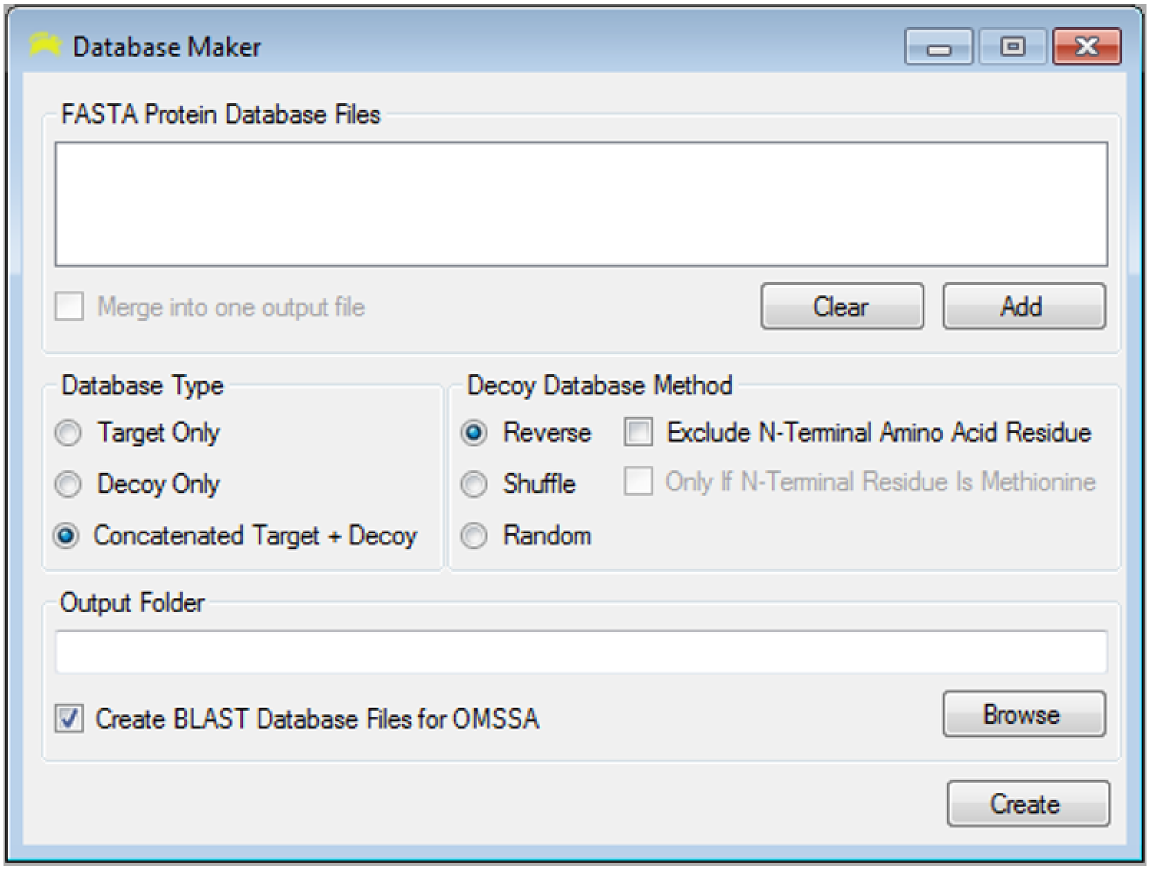


Figure 4.1: Analysis workflow of COMPASS. Database Maker generates BLAST-formatted protein databases for OMSSA. DTA Generator converts raw instrument data to text files for searching with OMSSA. FDR Optimizer performs FDR analysis at the spectrum/peptide level, followed by protein parsimony and FDR analysis at the protein level with Protein Herder. For quantitation, the workflow is supplemented by TagQuant, which performs spectrum/peptide-level quantitation, and Protein TagQuant, which performs protein-level quantitation. Since publication, Protein Herder and Protein TagQuant have been combined into a new program called Protein Hoarder.

ease future maintenance. The following sections summarize the different parts of COMPASS and the improvements made to them since the initial publication.

Database Maker. `Database Maker` creates protein databases for target-decoy searching of MS/MS spectra. Text files containing each protein sequence, in the FASTA format, are converted to a decoy version of the same length by reversing, shuffling, or generating random amino acids.[3–5] The decoy sequences are then concatenated to the input file and exported to another FASTA file. Additionally, protein sequences can be converted to the basic local alignment search tool (BLAST) format for use with OMSSA.[6] The GUI portion of the program (Figure 4.2) has been restructured to enable multiple database files at the same time. Internally, `Database Maker` now uses the `makeblastdb` program to generate BLAST databases instead of the now depreciated `formatdb`, both of which are provided by the National Center for Biotechnology Information (NCBI).

DTA Generator. The second program in the COMPASS workflow is `DTA Generator`, which reduces LC-MS/MS spectra data to `.txt` files for database searching (Figure 4.3). Various peak cleaning algorithms are used to simplify spectral data prior to searching; these include removal of unreacted precursors, electron-transfer dissociation (ETD) pre-processing to remove precursors, charge-reduced precursors,



Database Maker

FASTA Protein Database Files

□ Merge into one output file

Database Type

○ Target Only

○ Decoy Only

● Concatenated Target + Decoy

Decoy Database Method

● Reverse □ Exclude N-Terminal Amino Acid Residue

○ Shuffle □ Only If N-Terminal Residue Is Methionine

○ Random

Output Folder

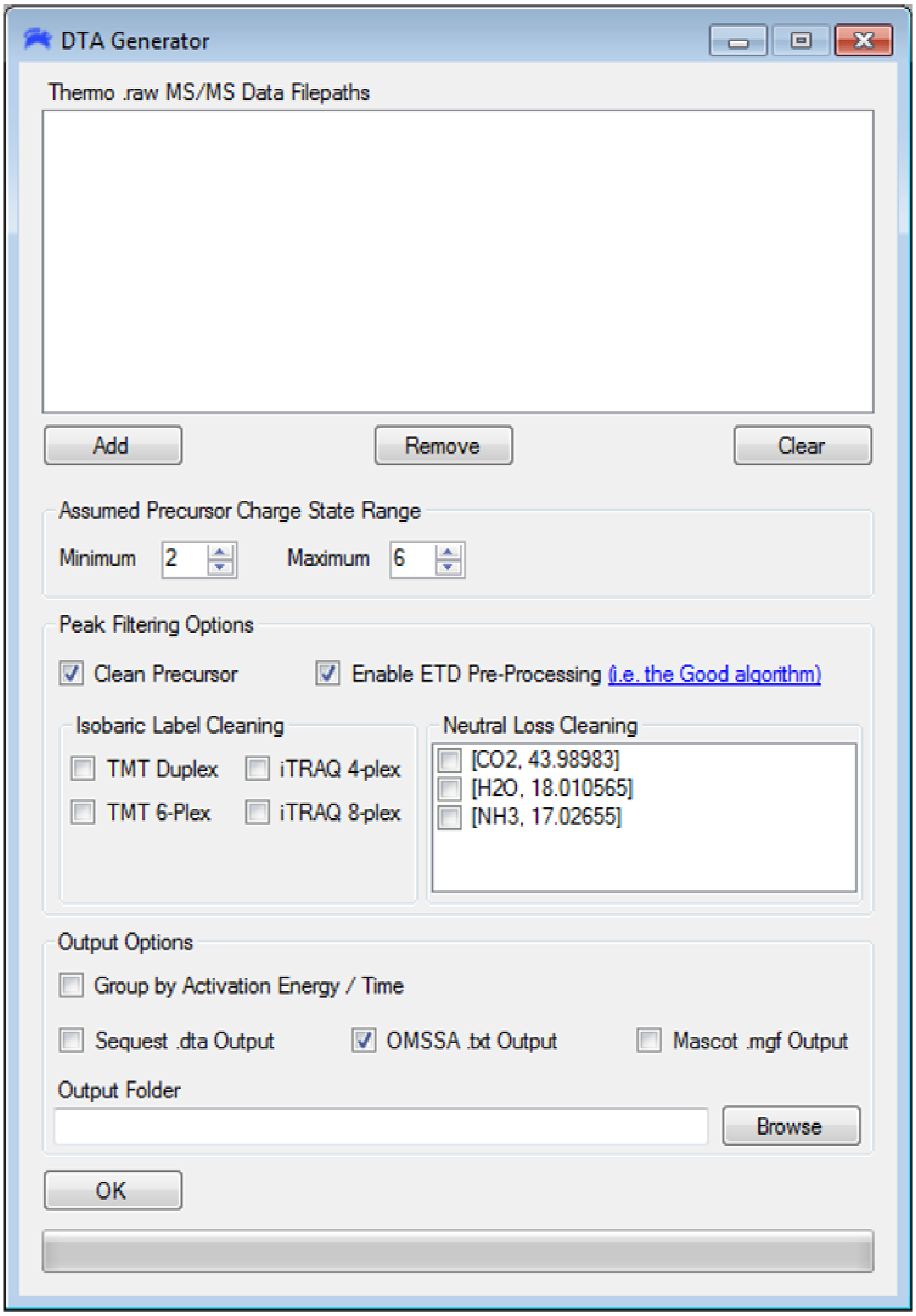
☑ Create BLAST Database Files for OMSSA

Figure 4.2: Database Maker. The GUI program used to manage, construct, and modify protein databases in the FASTA format. It is capable of constructing various forms of decoy databases used for false discovery analysis. An optional BLAST database can be produced for compatibility with OMSSA searching.

and neutral losses from charge-reduced precursors.[7] The outputs generated by the software are also usable by several other search algorithms. Although OMSSA is the focus, individual `.dta` files for SEQUEST or `.mgf` files for MASCOT are possible outputs.[2,8]

Since the initial publication, additional spectral filters have been added to allow the user more freedom in how the spectra are processed. These include specifying neutral loss products and isobaric labels for cleaning. The most significant improvement to `DTA Generator` since publication was a dramatic decreased in execution time (~50X). This was accomplished by converting the code to utilize multiple processor threads, as well as algorithmic improvements to spectral cleaning.

Open Mass Spectrometry Search Algorithm. `OMSSA` is a database search algorithm for proteomic datasets developed at the NCBI by Lewis Geer.[1] It uses a probabilistic scoring to associate a specific peptide sequence to an experimental spectrum. The program assigns an expectation value (e-value) to each peptide spectrum match (PSM) generated, stating the probability of matching that sequence to the spectrum by random chance. The smaller the e-value, the higher the confidence that peptide sequence produced the MS/MS spectrum. Further statistical analysis is performed in the `FDR Optimizer` program, discussed below. OMSSA provides the option to produce a `.csv` output of all the PSMs generated. This format is the



Thermo .raw MS/MS Data Filepaths

Assumed Precursor Charge State Range

Minimum 2 Maximum 6

Peak Filtering Options

☑ Clean Precursor ☑ Enable ETD Pre-Processing

Isobaric Label Cleaning

□ TMT Duplex □ iTRAQ 4-plex

□ TMT 6-Plex □ iTRAQ 8-plex

Neutral Loss Cleaning

□ [CO2, 43.98983]

□ [H2O, 18.010565]

□ [NH3, 17.02655]

Output Options

□ Group by Activation Energy / Time

□ Sequest .dta Output ☑ OMSSA .txt Output □ Mascot .mgf Output

Output Folder

Figure 4.3: DTA Generator. The program takes spectral data in Thermo's `.raw` format and generates a `.txt` of the processed spectra. Processing includes removing peaks that do not provide sequence-informative results (i.e., neutral loss). The program is capable of producing outputs for OMSSA, SEQUEST and Mascot search algorithms.

basis for all the other programs in COMPASS. It is easily opened and manipulated by spreadsheet programs (e.g., Microsoft Excel) and is human readable. This is in contrast to a majority of other proteomic software, where `.xml` or a proprietary format are used. Those formats make modifying and parsing the data more difficult than a `.csv` file.

High Throughput Condor for OMSSA. Arguably the biggest improvement to COMPASS since publication is the addition of the High Throughput Condor (HTCondor) system for improving OMSSA searching times. In brief, HTCondor is a computational management system for scheduling processing jobs across a distributed network of computers. Computers voluntary join a HTCondor network which enables them to donate their free CPU cycles to other processes, increasing the overall processing power of the network. This is ideal for large universities, where there is a large number of computers on a common network, and a majority of those computers (e.g., computer labs, servers, kiosks, office computers) are not in use twenty-four-seven. The HTCondor system intelligently monitors CPU activity on each attached computer, and given a certain amount of inactivity, reassigns its CPU to process jobs waiting in a global queue. If HTCondor detects new local activity on that computer (e.g., keyboard or mouse movement, a local processing job, etc.), it will either pause the global job, or automatically transfers it to another inactive computer. Given the large size of the HTCondor network on the University of Wisconsin-Madison campus (~7,000 CPUs) there is a high probability that there will always be multiple computers available for analysis. This large network provides million of CPU hours to researches all across campus. From the HTCondor website, they state that ”from July 2011 to June 2012, the [Center for High Throughput Computing] provided 70 Million CPU hours to campus researchers and off-campus collaborators.”

Shortly after COMPASS was published, I developed a GUI program called `Coondornator` to provide a method for searching MS/MS data *via* OMSSA over the University of Wisconsin-Madison's HTCondor network. The program first transfers `.dta` files (generated by `DTA Generator`) from a user's computer to the Coon Group computer cluster, which is its own 17-CPU HTCondor network. If these computers are idle, they automatically start processing each submitted OMSSA job. If more than 17 OMSSA searches are submitted at once, overflow jobs are automatically routed to the Center of High Throughput Computer (CHTC) HTCondor cluster (~6,800 CPUS) for analysis. It is common to have 50-100 OMSSA searches going at any give time. When the OMSSA searches are completed, the resulting `.csv` file containing all the PSMs are transferred back to the Coon Group computer cluster for storage. This program provides a seamless integration between the HTCondor network and an end user's computer, making high throughput computing no harder than running a program on their computer. The Coon Group routinely searches thousands of `.dta` files containing million of spectra on the HTCondor network each week.

Previously, users would search all the MS/MS spectra from a single LC-MS/MS experiment on their desktop computers. Now with `Coondornator` and HTCondor, a single LC-MS/MS experiment is broken up into smaller sets of spectra (e.g., 1000 spectra per set), searched individually, and then recombined when all the searches are complete. This represents a significant throughput gain compared to searching files on individual desktop computers. With execution times decreasing on average about 30-50X. For example, it would take about 30-40 minutes on a desktop computer to search all the MS/MS spectra from a one hour LC-MS/MS experiment of a tryptic digestion of whole-cell yeast cells. In contrast, if all the MS/MS spectra were split into groups of 1000 spectra each, and searched using `Coondornator` over the distributed HTCondor network, the same results could be generated in ~1 minute, a 30X decrease in execution time.

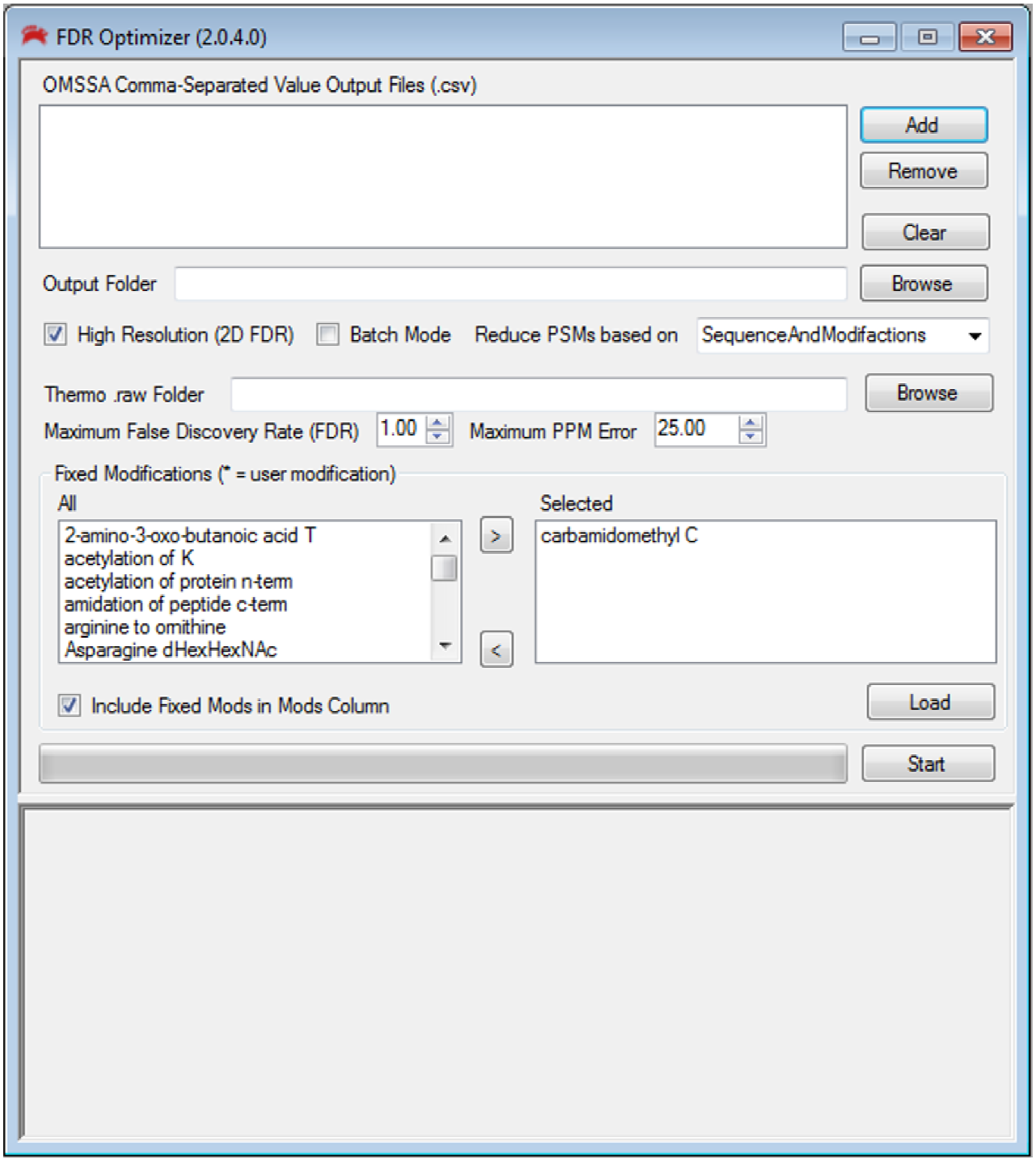
FDR Optimizer. `FDR Optimizer` filters PSMs generated from `OMSSA` to control for false identifications (Figure 4.4). The program maximizes the number of true positive identifications at a given false discovery rate (FDR), typically set to under

1%. FDR Optimizer can work on either low-resolution MS data with a simple e-value filter, or on high-resolution MS data with a two dimensional filter on precursor mass error and e-value. To use FDR Optimizer, both target and decoy protein sequences have to be searched with OMSSA on the same set of spectra for adequate false discovery filtering.

For low-resolution datasets, PSMs are first loaded into the program and the best scoring PSM (i.e., lowest e-value) for each spectrum is saved and all other PSMs are discarded. The remaining PSMs are then sorted on their e-value, from smallest to largest. Each PSM also has a flag to indicate whether it resulted from a target protein or a decoy protein. A counter for both the number of targets (T) and decoy (D) peptides identified is kept as the program iterates over the sorted PSMs. When the false discovery rate (Equation 4.1) increases over some specified value (e.g., 1%), the program stops the iteration.

$$\mathsf{FDR} = \frac{\mathsf{D}}{\mathsf{T} + \mathsf{D}} \tag{4.1}$$

The PSMs that represent the true identifications and which have already been processed are then outputted to a `.csv` file. Known decoy peptides that pass this filter are exported to a decoy-specific `.csv` file that is used later for protein-level FDR analysis.



FDR Optimizer (2.0.4.0)

OMSSA Comma-Separated Value Output Files (.csv)

Output Folder

☑ High Resolution (2D FDR) □ Batch Mode Reduce PSMs based on SequenceAndModifactions

Thermo .raw Folder

Maximum False Discovery Rate (FDR) 1.00 Maximum PPM Error 25.00

Fixed Modifications (* = user modification)

All

2-amino-3-oxo-butanoic acid T
acetylation of K
acetylation of protein n-term
amidation of peptide c-term
arginine to omithine
Asparagine dHexHexNAc

Selected

carbamidomethyl C

☑ Include Fixed Mods in Mods Column

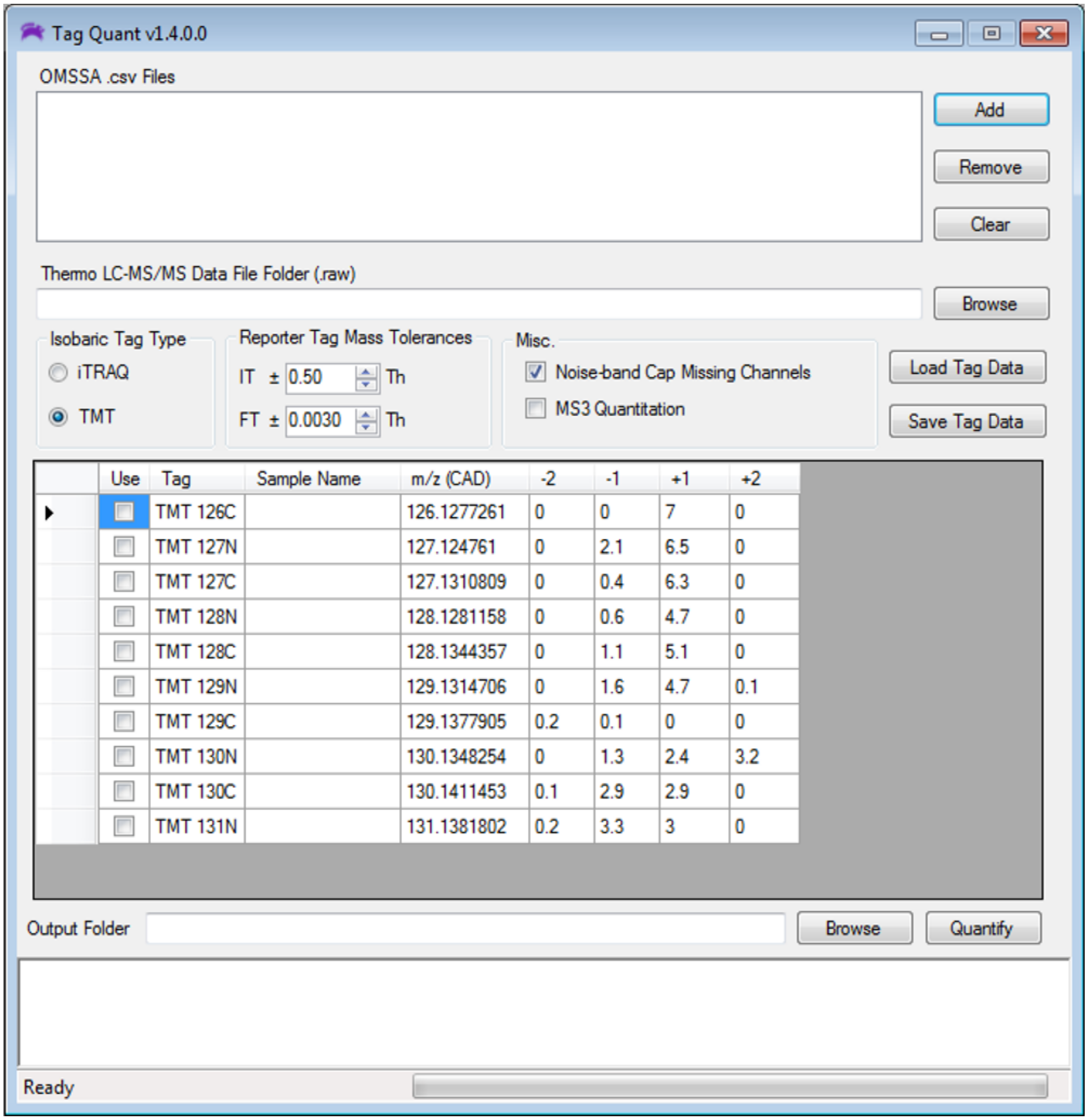
Figure 4.4: FDR Optimizer. The new GUI for `FDR Optimizer`, this version combines all four versions into one program for a simplified user experience.

High-resolution MS datasets can be processed with an additional filter to increase the number of identifications. First each PSM is read into the program and its precursor mass error is determined from the MS spectrum that triggered the MS/MS event. The median precursor mass error of all PSMs is then computed and each PSM is corrected by this value. This process corrects any systematic mass error the mass spectrometer had, and usually reduces the mass errors to <5 ppm. `FDR Optimizer` then iteratively sets a maximal ppm error allowed (i.e., from 1 to 100 ppm), and filters the PSMs to contain only precursor mass errors lower than the maximum. These filtered PSMs are then processed identically to the low-resolution analysis described above. This whole process is then repeated with a slightly larger maximal ppm error, and the number of identifications is recorded. The program tries all possible maximal ppm errors and reports the ppm error that produced the most true identifications at the end. This maximizing algorithm increases the number of true identifications produced over the simple low-resolution filter by roughly 10-15%.

Since publication, `FDR Optimizer` has been completely rewritten. Previously, four separate programs were used and maintained: `Low-Resolution FDR Optimizer`, `FDR Optimizer`, `Batch Low-Resolution FDR Optimizer`, and `Batch FDR Optimizer`. The current version simplifies the user experience by combining all four programs into one, with simple option check-boxes to indicate the desired analysis (Figure 4.4). Improvements to the FDR analysis and maximizing algorithms have also lead to large decreases in execution times (~10-20X).

TagQuant. The `TaqQuant` program extracts and processes isobaric labeling quantitative information from MS/MS spectra (Figure 4.5). It is compatible with both common types of isobaric labels, TMT and iTRAQ.[9,10] `TagQuant` obtains intensities of the reporter ions of interest from the raw data. These intensity values are subsequently denormalized by multiplying by the ion injection time to yield the number of ion counts detected, a quantity which can be fairly compared across different spectra and analyses. Purity correction is then applied using user-specified purity data provided by the manufacturer.[11] Finally, normalization is performed such that the total intensity of each tag is equal, accounting for differences that arise when samples are mixed.

Numerous improvements have been made to the publication version of `TaqQuant`. With the advent of high-resolution TMT tags, where two quantiation channels are separated by a very small mass difference (6.32 mDa), additional logic had to be added to handle it.[12,13] Users also started to mix and match channels between different manufacturing lots, resulting in non-standard purity values. This, and other issues, were corrected by providing the user full control over which labels they



OMSSA .csv Files

Thermo LC-MS/MS Data File Folder (.raw)

Isobaric Tag Type

○ iTRAQ

● TMT

Reporter Tag Mass Tolerances

IT ± 0.50 Th

FT ± 0.0030 Th

Misc.

☑ Noise-band Cap Missing Channels

□ MS3 Quantitation

Use	Tag	Sample Name	m/z (CAD)	-2	-1	+1	+2
□	TMT 126C		126.1277261	0	0	7	0
□	TMT 127N		127.124761	0	2.1	6.5	0
□	TMT 127C		127.1310809	0	0.4	6.3	0
□	TMT 128N		128.1281158	0	0.6	4.7	0
□	TMT 128C		128.1344357	0	1.1	5.1	0
□	TMT 129N		129.1314706	0	1.6	4.7	0.1
□	TMT 129C		129.1377905	0.2	0.1	0	0
□	TMT 130N		130.1348254	0	1.3	2.4	3.2
□	TMT 130C		130.1411453	0.1	2.9	2.9	0
□	TMT 131N		131.1381802	0.2	3.3	3	0

Output Folder

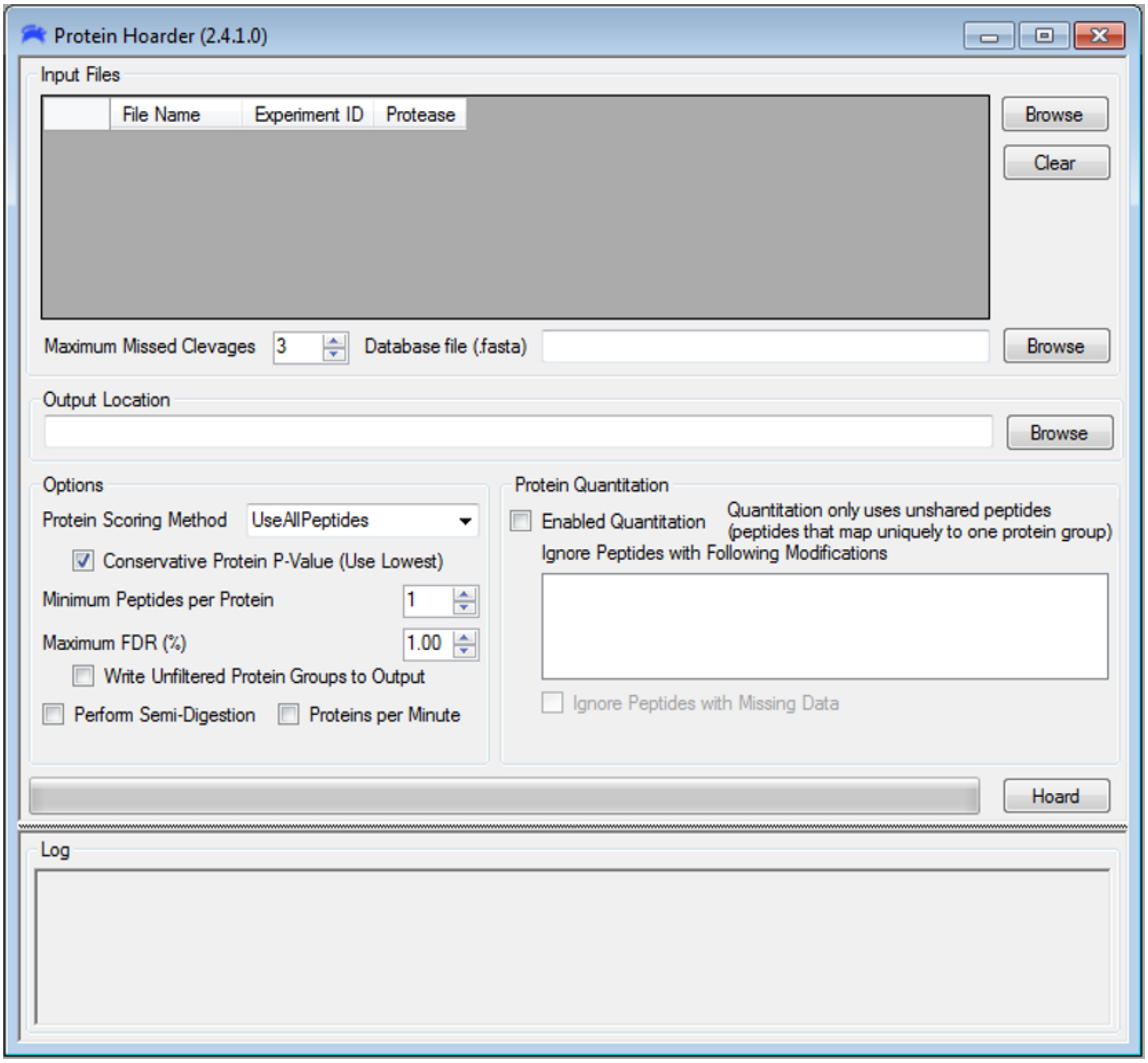
Ready

Figure 4.5: TagQuant. The updated GUI for `TagQuant` allows users to specify the exact labels used and their relative purities. Options for noise-band capping missing channel and quantifying from an MS^3 scan are also included.

used to quantify and their respective purities. This change also enables `TagQuant` to handle any type of isobaric label later developed without making changes to the program itself.

Another heavily used feature that was added was capping missing channels with the noise-band intensity. Sometimes a peak at an expected isobaric tag *m/z* is not present in the MS/MS spectra. Before, this missing value was set to 0, which would drastically distort the ratio between quantitation channels. In the updated version, `TagQuant` assigns the missing value to the noise level at the *m/z*. This conservative approach mitigates the distortion of ratios between quantitative channels. An additional feature that was added was enabling quantitation from MS^3 spectra. This is the result of purification methods for improving the interference problem of isobaric labels, such as QuantMode and the MS^3 methods.[14,15]

Protein Hoarder. `Protein Hoarder` infers the most likely proteins identified based on the peptides validated by `FDR Optimizer` (Figure 4.6). The program was initially called `Protein Herder`, but the program was completely rewritten after publication, and the name was changed to indicate that it is a new program. In this program, peptides are assembled into protein groups based on the law of parsimony, i.e., minimizing the number of protein groups while accounting for all the identified peptides. False discovery analysis is also performed at the protein-level, along with



Maximum Missed Clevages 3

Database file (.fasta)

Protein Scoring Method UseAllPeptides

☑ Conservative Protein P-Value (Use Lowest)

Minimum Peptides per Protein 1

Maximum FDR (%) 1.00

□ Write Unfiltered Protein Groups to Output

□ Perform Semi-Digestion □ Proteins per Minute

□ Enabled Quantitation

Quantitation only uses unshared peptides (peptides that map uniquely to one protein group)

Ignore Peptides with Following Modifications

□ Ignore Peptides with Missing Data

Figure 4.6: Protein Hoarder. The new version of `Protein Herder` which assembles peptide identifications into protein groups. This version also performs protein quantitation at the same time the protien groups are being assembled.

quantitative analysis if requested. The outputs of the program include a `.csv` file containing all the identified protein groups, along with which peptides are mapped to the groups.

The biggest change between the original and current version is how the peptides are found within the candidate proteins. Previously, each peptide sequence was searched against the whole protein database using brute force. For large databases such as the human proteome, the number of proteins could reach over 150,000 when isoforms and both target and decoy proteins are considered. Even if only 20,000 PSMs were identified, that means 30 billion string comparisons must be made (20,000 x 150,000). This made the original program very slow, and could take up to half a day to assemble protein groups for a human sample. The new algorithm forgoes the string search and uses enzymatic cleavage of the proteins to find the associated peptides. The program preforms an *in silico* digestion of all the proteins and if a generated peptide matches one of the input PSMs, that protein is saved. This process greatly speeds up the whole program, and the same human sample that took half a day to assemble now takes less than 2 minutes.

Assembled protein groups are further filtered for false discovery using a similar method to `FDR Optimizer`. Here, the p-value of the protein group (which is the product of all the peptide's e-values) is the ordering metric and the groups are filtered to a specified FDR (e.g., 1%). In the publication version of COMPASS, there was another program that handled protein quantitation (`Protein TagQuant`) by summing up all the peptide quantitation for a individual protein group. Peptides that are not shared between protein groups (i.e., a unique peptide to the protein group) have their quantitation summed and reported for the protein group. This program was embedded into `Protein Hoarder` since all the required information for quantitation was already present in the program. This removed the need for using `Protein TagQuant` altogether, and it was removed from COMPASS.

LoToR. The final program in COMPASS was not present in the initial version. This program is called `LoToR` (Localize To Residue) and improves the localization of post translational modifications to specific residues on peptides and proteins. Although `OMSSA`, and other search algorithms, are capable of identifying modification events on peptides, it often does not place the PTM on the correct residue. To address this, `LoToR` was created to add more rigorous statistical power in localizing PTMs to specific sites.

`LoToR` is uses the AScore algorithm as the primarily metric for assigning statistical confidence.[16] In brief, for each PSM that contains a PTM, all possible peptide isoforms are generated. Each of these isoforms represents the unique combination of PTMs applied to every possible site on the peptide. Then each isoform is fragmented *in silico* and matched to the MS/MS spectrum. After matching, each isoform is compared to every other isoform generated, and the set of fragments that can distinguish them apart are called ’site-determining fragments’ (SDFs). The number of identified SDFs for each isoform is then compared, and the two isoforms that have the biggest difference in the number of identified SDFs is declared the best possible isoform. The AScore for this pair of isoforms is then computed, and if the value is above some defined value (typically 13), it is declared localized. LoToR is capable of handling any modification (e.g., phosphorylation, acetylation, ubiquitination, etc.).

CSMSL: C# Mass Spectrometry Library

Mass spectrometry-based proteomics is a relatively young field that is rapidly evolving and new techniques and technologies are consistently being developed, prompting the need for custom software tools to analyze the data. There are typically three ways to analyze a proteomic dataset: 1) process it through a full-fledged GUI program that has already been developed, 2) manually process the data, or 3) extract data with software tools and analyze with other software (e.g., Microsoft Excel). Often a mixture of these three processes are needed to fully analyze a dataset. However, sometimes a complete GUI program is not available for a specific type of data analysis, or, due to the complexity of the data, manual analysis of a dataset in Excel or through a spectrum browser (e.g., Thermo XCalibur) is a daunting and time-consuming task. These situations are ideal for a custom analysis program that could facilitate the analysis. Unfortunately, creating custom programs is not straightforward: 1) not every researcher knows how to program, and 2) there isn't a free, simple programming environment for accessing and manipulating such complex data. The first problem is not easily addressed, but the second one is. Although there are many tools available for MS analysis on the internet, most are difficult for novice programmers and are challenging to adapt to a specific need. To fill the gap, I have designed a large proteomic programming library to simplify the data management and manipulation of large-scale proteomic data. It is written for Windows using the .NET Framework V4.0 in the C# programming language. It is called C# Mass Spectrometry Library (CSMSL) and is freely available at `https://github.com/dbaileychess/CSMSL`.

The goals of CSMSL are to provide an easy-to-use, powerful, feature-rich library of .NET C# objects and methods to enable even novice programmers the ability to analyze proteomic data quickly. Simplicity is key. Calculating the mass of the peptide sequence 'CSMSL' only requires the following two lines:

```
Peptide peptideA = new Peptide("CSMSL");
```

```
Console.WriteLine(peptideA.MonoisotopicMass);
// outputs : 539.20835516707
```

In addition to simple syntax, CSMSL is designed with performance in mind, allowing even computationally intensive calculations to be completed quickly. For example, a complete yeast database (6,627 proteins) can be loaded from a FASTA file, digested with trypsin (up to 3 missed cleavages, 5 to 35 amino acids in length) in under 2 seconds. If the calculation for the $[M+H]^+$ *m/z* of each of the 913,740 resulting peptides is included, the total time only goes up to 4 seconds (this includes full chemical formula determination). While CSMSL is not expected to meet the performance of advanced compiled languages (e.g., C/C++, Fortran, etc.), its adequate performance plus simplicity of use are sure to be helpful in analyzing data in new and creative ways without significant overhead.

The following sections will succinctly 1) describe the design of the library, 2) show a few example code segments indicating its use, and 3) highlight various features and abilities of the library.

CSMSL Design. The CSMSL package is divided into three projects. The main project is the library itself (`CSMSL.csproj`) which contains all the code and objects to program with. This project will be described in greater detail in the sections to follow. The other two projects are primarily used for teaching and development purposes, and will be described here.

The teaching project is `CSMSL.Examples.csproj`, which contains short segments of code to show the intended use of the library. It is written to aid novice programmers in learning how to program better and how to use the library. It covers a series of example code to demonstrate how to create peptides and proteins, digest proteins, fragment peptides, read in spectral data, among many others. This project is completely separate from CSMSL and is only used to demonstrate the features of the main library.

The development project is `CSMSL.Tests.csproj`, which hosts all the unit tests for the library. A unit test is a short piece of code that tests one, and only one aspect of the library, hence the term 'unit'. In brief, a short segment of code is written to preform some action (e.g., digest a protein), and the final line contains an assertion statement, declaring that some value needs to possess some trait (e.g., that 5 peptides are produced from a digestion of a certain protein). These assertions can be as simple as an equality ($\mathtt{numOfPeptides} == 5$), a comparative condition ($\mathtt{numOfPeptides} < 5$), or much more advanced comparisons. Regardless, the point of unit tests is to provide fine grain support and testing for the main project. If any source code is added to the project or a portion of code is changed, all the unit tests report back to the developer if their one piece of functionality is still producing the same result. If not, the developer has a good idea of where the new bug is introduced as only certain unit tests will fail. This helps ensures that new features do not affect other parts of the library or produce unintended bugs. CSMSL is heavily tested, especially on the components that are most commonly used.

There also exists a handful of other projects that supply support for third-party tools and access to raw spectral data from different MS vendors. These are located under the `CSMSL/IO` directory and can be added to a project when needed. The projects that support reading in raw spectral data will be discussed in the features section below. Since CSMSL has been developed, it has been heavily incorporated into the source code of COMPASS. This helps simplify every program within COMPASS, as many redundant sections of code were replaced by objects from CSMSL. It also helps speed up bug fixes and feature additions, as all programs that use CSMSL will benefit from improvements in its code.

CSMSL Examples. Functional coding examples are a great way to dive into any programming language/library. CSMSL provides a number of example programs, contained within the `CSMSL.Examples` project, so that people can learn the tools and experiment with different features. Below are a series of examples showing the simplicity and power CSMSL offers. All of the examples are written in the

C# language and should be straightforward enough that even non-programmers should be able to follow them.

We will start with the most basic, but most commonly used features: proteins and peptides. The following code first constructs a new peptide object in memory, labels it as ’peptideA’ and then prints its monoisotopic mass to the console window.

```
Peptide peptideA = new Peptide("FLTTSNALKEN");
Console.Write(peptideA.MonoisotopicMass);
// outputs: 1236.635016661
```

Of course there are a plethora of tools and websites that could calculate the monoisotopic mass of a peptide sequence, but the novel aspect is its simplicity and the ability to programmatically control it.

Peptides and proteins can be modified post transitionally and CSMSL enables easy methods for modifying peptides. Taking the previous example further, to modified the serine residue (’S’) with a phosphorylation is easy:

```
Peptide peptideA = new Peptide("FLTTSNALKEN");
ChemicalFormula phospho = new ChemicalFormula("HPO3");
peptideA.SetModification(phospho, 'S');
Console.Write(peptideA.MonoisotopicMass);
// outputs: 1316.60134718175
```

Only two new lines are inserted. Line 2 creates the phosphorylation modification (labeled as 'phospho'), and introduces another CSMSL object called ChemicalFormula, which represents a chemical structure. The third line sets the 'phospho' chemical formula to modified all the serine residues in peptideA. Since there is only one serine in peptideA, only one phosphorylation is added, resulting in a mass of 1316.601347.

Another important feature often used is protein digestion. In nature, peptides often arise from the proteolyic digestion of intact proteins by proteases, such as trypsin. CSMSL can do the same thing *in silico* that is done in a test tube. Below is an example of a tryptic digestion of a single protein. It produces a list of Peptide objects which are then printed to the screen.

```
Protein proteinA = new Protein("MMGFKQLITTGSSSRSSSSKDTSST");
List<Peptide> peptides = proteinA.Digest(Protease.Trypsin);
Console.Write(peptides);
// outputs: MMGFK, QLITTGSSS, SSSSK, etc...
```

The first line creates a new protein object in memory, just like the peptide example above. The second line takes the created protein (proteinA) and performs a digestion with trypsin. The result (the left hand side of the equation) is a list of Peptide objects. The final line then takes all those peptide sequences and prints them to the screen. This is a simple digestion, but there exist many more options, such as maximum and minimum peptide length, max missed cleavages, partial digestion, etc. All these options will not be explored here, but almost anything you could do on a physical protein/peptide, can be performed *in silico* using CSMSL.

Full Proteome Tryptic Digestion Example. The code section shows a complete example of a tryptic digestion of a yeast proteome. Its inclusion here is to demonstrate that a fairly complicated task can be performed with only a few lines of code that should be easily understandable to a novice programmer.

```
using (FastaReader reader = new FastaReader("yeast.fasta"))
{
 Protease trypsin = Proteases.Trypsin;
 int max = 3;  // Maximum number of missed cleavages
 foreach (Protein protein in reader.ReadNextProtein())
 {
  foreach (Peptide peptide in protein.Digest(trypsin, max))
  {
    Console.WriteLine(peptide.MonoisotopicMass);
  }
 }
}
```

Line 1 opens up a connection to a protein database file named ”yeast.fasta” located on the computer. The `FastaReader` object provides methods for reading and accessing proteins contained in a FASTA-formatted file. The third line sets up a variable that represents the typsin protease. Line 4 sets another variable defining the maximum number of missed cleavages allowed for the digestion. The fifth line iterates over each `Protein` within the FASTA file, by calling the `ReadNextProtein()` method. The lines 6 through 11 are then performed for each protein read in by line 5. Line 7 iterates over every peptide generated from the digestion of the protein by trypsin and a maximum missed cleavage of three. Again, lines 8 through 10 are repeated for every generated peptide. The last important line is line 9, where the peptide’s monoisotopic mass is written to the console window on the computer. In only 12 lines of code, a complicated task is accomplished using CSMSL objects and methods. Similarly, other proteomic analysis and computations can be expediently coded and performed. These examples can be found in more details in the `CSMSL.Examples.csproj` project file.

CSMSL Features and Objects. The CSMSL library has too many features to list in full, so only a few of the most important features will be highlighted here. Since this is primarily a proteomic library, it will start off with proteins and chemicals and then transition to spectral classes.

Starting from the smallest object and growing bigger, elemental isotopes represent the basic building block of everything else that has mass. Each isotope has a few intrinsic properties, most importantly its mass and the element it belongs to. A single element may contain a set of different isotopes (e.g., ^{12}C, ^{13}C, and ^{14}C), and the naturally most-abundant isotope is declared the principal isotope of the element. Thus elements and their most abundant isotopes are interchangeable with each other (i.e., ^{12}C and C refer to the same object). When other isotopes are needed, you need to specify which isotope you want to use (e.g., C{13} mean you want ^{13}C instead of ^{12}C). This feature is important because stable isotope quantitative labeling is a very common analysis, and I wanted to design the library with it in mind. All the elements, and thus isotopes, are assembled into the periodic table of elements for easy access.

In CSMSL, chemical formulas are represented as a set of isotopes without any spatial connectivity. Keeping the three dimensional structure of molecules is not an important aspect of most proteomic work, and I purposefully left this out in favor of speed and memory savings. Additionally, a chemical formula generator is available to list all possible chemical formulas when given an exact mass. Such features are used to indentify unknown peaks in high-resolution mass spectra. The mass of a chemical formula is the simple summation of all of its isotopes. Almost every other object in this library is a chemical formula (e.g., proteins, peptides, amino acids, modifications, etc.) with additional properties of its own. Amino acids are simply a chemical formula with a character symbol to represent which one it is. The 20 common amino acids are prebuilt by the library and ready to use, but custom amino acids can be added easily.

Probably the most important classes in the library are the protein and peptide classes. Since both a protein and peptide can be thought of as a string of amino acids, both classes are modeled off a single base class called `AminoAcidPolymer`. This class can be thought of as an fancy array of amino acids, spanning from the N-terminus to the C-terminus. Each location on this array (i.e., amino acids or termini) can be modified by a chemical formula. The mass of the `AminoAcidPolymer` is again the summation of all its amino acids and modifications. Peptides have special methods for producing fragments ions (e.g., a, b, c, x, y, z-type, as well as others). Fragment ions are also chemical formulas, but they keep track of what amino acids and modifications are contain in each fragment. This is particularly useful when matching fragment ions against a mass spectrum, as it fully annotates the spectrum during the matching steps. Proteins have methods for proteolytic digestion by any enzyme. Users can also define their own proteases to use, and they will be compatible with all the features of digestion (e.g., missed cleavages, min/max lengths, semi-digestions, etc.).

Finally, a set of spectral-related classes are included to provide easy access to spectral data. In CSMSL spectra are comprised of a ordered list of `Peaks`, which represents the *m/z* and intensity of the peak. These `Peaks` objects can be extended to contain additional information (e.g., charge, noise, baseline intensity, etc.) provided by the instrument. The `Spectrum` class contains the data of a specific spectrum and provides a series of methods for easy data access. Since a large part of proteomic analysis is looking for peaks within a spectrum, probably the most important method is the `GetClosestPeak()` method. This uses a binary-search algorithm to quickly find the closest peak to a given *m/z* value, and this operation takes on average $O(\log N)$ time, where N is the number of peaks in the spectrum. Even with complex spectrum of 1,000 peaks, it only takes about 7 comparisons to locate a peak.

Spectral Data Access. A very useful feature CSMSL provides is access to raw spectral data collected by the MS. Instrument vendors usually offer an application programming interface (API) for accessing data from their propriety data formats (e.g., `.raw` for Thermo, `.d` for Agilent, etc.). While it is possible to use them on own, a few factors make them difficult to implement. First, they are geared to more advanced programmers and often have incomplete documentation. This makes learning how to access the data difficult, even for good programmers. For people who don't know how to program at all, it would be very difficult to understand and make work. Secondly, each instrument vendor creates their own API which is incompatible with everyone else's. Thus if you desire to analyze two different types of data with your program, you'll have to use both APIs to achieve the same result. This additional code can often lead to bugs and frustration. Lastly, you have to be an expert in each API in order to fully use their capabilities. CSMSL solves these issues by having a single and simple interface for accessing the data, no matter where or how the data was produced.

The following example shows how to read in every MS spectra from a `.raw` file generated from a Thermo mass spectrometer.

```
MSDataFile dataFile = new ThermoRawFile("somerawfile.raw")
dataFile.Open();
foreach (MSDataScan scan in dataFile)
{
    Console.WriteLine('Number of Peaks: '+scan.PeakCount);
}
\\ outputs:
\\   Number of Peaks: 1052
```

```
9  \\    Number of Peaks: 523
10 \\    etc..
```

First, in line 1 a mass spectrum data file is constructed from a file on the computer, named ”somerawfile.raw”. The second line opens a connection to the file, and the third line iterates over each MS scan within that file. The number of peaks contained within that scan is then printed to the screen. The beauty of this example is that the first line could be changed to:

```
1 //MSDataFile dataFile = new ThermoRawFile("somerawfile.raw")
2 MSDataFile dataFile = new AgilentDDirectory("somerawfile.d")
```

and the program would continue to work, even though it is now accessing MS data collected by an Agilent mass spectrometer. Having this sort of flexibility built in from the start enables programmers to program their tools once and have it work with data from any source supported. As of this publication, only Thermo `.raw`, Agilent `.d` and `.mzML` formats are supported. However, other vendors could be added in the future without breaking code. While there are too many features to be fully explained here, the concept of a simple and consistent way to access spectral data is a key component of CSMSL.

References

[1] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, “Open mass spectrometry search algorithm,” *J Proteome Res*, vol. 3, no. 5, pp. 958–64, 2004.

[2] J. K. Eng, A. L. McCormack, and J. R. Yates, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *J Am Soc Mass Spectrom*, vol. 5, no. 11, pp. 976–89, 1994.

[3] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” *Nat Methods*, vol. 4, no. 3, pp. 207–14, 2007.

[4] L. Blanco, J. A. Mead, and C. Bessant, “Comparison of novel decoy database designs for optimizing protein identification searches using abrf sprg2006 standard ms/ms data sets,” *J Proteome Res*, vol. 8, no. 4, pp. 1782–91, 2009.

[5] R. E. Moore, M. K. Young, and T. D. Lee, “Qscore: an algorithm for evaluating sequest database search results,” *J Am Soc Mass Spectrom*, vol. 13, no. 4, pp. 378–86, 2002.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J Mol Biol*, vol. 215, no. 3, pp. 403–10, 1990.

[7] D. M. Good, C. D. Wenger, G. C. McAlister, D. L. Bai, D. F. Hunt, and J. J. Coon, “Post-acquisition etd spectral processing for increased peptide identifications,” *J Am Soc Mass Spectrom*, vol. 20, no. 8, pp. 1435–40, 2009.

[8] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, vol. 20, no. 18, pp. 3551–67, 1999.

[9] A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed, and C. Hamon, “Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms,” *Anal Chem*, vol. 75, no. 8, pp. 1895–904, 2003.

[10] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin, “Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents,” *Mol Cell Proteomics*, vol. 3, no. 12, pp. 1154–69, 2004.

[11] I. P. Shadforth, T. P. Dunkley, K. S. Lilley, and C. Bessant, "i-tracker: for quantitative proteomics using itraq," *BMC Genomics*, vol. 6, p. 145, 2005.

[12] G. C. McAlister, E. L. Huttlin, W. Haas, L. Ting, M. P. Jedrychowski, J. C. Rogers, K. Kuhn, I. Pike, R. A. Grothe, J. D. Blethrow, and S. P. Gygi, "Increasing the multiplexing capacity of tmts using reporter ion isotopologues with isobaric masses," *Anal Chem*, vol. 84, no. 17, pp. 7469–78, 2012.

[13] T. Werner, I. Becher, G. Sweetman, C. Doce, M. M. Savitski, and M. Bantscheff, "High-resolution enabled tmt 8-plexing," *Anal Chem*, vol. 84, no. 16, pp. 7188–94, 2012.

[14] C. D. Wenger, M. V. Lee, A. S. Hebert, G. C. McAlister, D. H. Phanstiel, M. S. Westphall, and J. J. Coon, "Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging," *Nat Methods*, vol. 8, no. 11, pp. 933–5, 2011.

[15] L. Ting, R. Rad, S. P. Gygi, and W. Haas, "Ms3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics," *Nat Methods*, vol. 8, no. 11, pp. 937–40, 2011.

[16] S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush, and S. P. Gygi, "A probability-

based approach for high-throughput protein phosphorylation analysis and site localization," *Nat Biotechnol*, vol. 24, no. 10, pp. 1285–92, 2006.

Chapter 5

THE FUTURE OF INTELLIGENT DATA ACQUISITION METHODS

Summary

Intelligent data acquisition methods are technologies on the forefront of mass spectrometry. These methods utilize data analysis algorithms—typically performed after acquisition, during the acquisition of MS spectra to improve data quality. The ability to immediately analyze spectra and then make informed decisions on how to proceed is important in separation-based analyses, where an analyte is only accessible for a short time period. This document has looked at the history of MS acquisition methods and began with a discussion on data-dependent acquisition and other acquisition methods. The second chapter described our work on intelligent data acquisition (IDA) methods; we developed the first online spectral database search (*inSeq*) to improve multiple aspects of data acquisition. The following chapter continued on this theme and focused on improving the reproducibility of peptide identification by using real-time elution ordering scheduling. The fourth chapter took a behind-the-scenes look on the programming environments used and developed to enable IDA methods. In this chapter, various challenges that confront intelligent data acquisition—both scientific and practical, are discussed

and possible solutions are proposed.

Introduction

Data-acquisition methods are an important part of mass spectrometry analysis. This is especially true when the MS is coupled to a separation technique such as liquid or gas chromatography (LC-, GC-MS). Here a complex sample is separated over time and only a small subset is analyzed at any point in time. In LC-MS, analytes may only elute for 15-30 seconds and in GC-MS times are even shorter (5-10 seconds); in either case the MS has a limited amount of time to detect them. For proteomic work, to identify the analytes (e.g., peptides and proteins) a MS/MS spectra must also be collected to determine the amino acid sequence. Thus it is not only important to detect the precusor in a MS scan, but it also has to be isolated, fragmented and mass analyzed again (MS/MS) to determine its identity, all of which takes time. Given the complexity of proteomic samples (thousands of proteins digested into hundreds of thousands of peptides) the main challenge becomes one of time-management, i.e., how should the mass spectrometer use its time?

The most straightforward answer is to speed things up, make the mass spectrometer faster and more sensitive so it can spend less time per analyte, and therefore gain access to more of the proteome. The fastest mass spectrometers can achieve nearly 20 Hz scan rates while still being sensitive and selective enough to identify peptides. But increased speed can only solve so much of the time-management problem. Take yeast for example, with approximately 6,600 proteins that produce half a million peptides when digested with trypsin (1 missed cleavage). How long would it take to sample each peptide? Assuming a 20 Hz acquisition rate, an 100% identification rate, and a perfect LC separation, it would take at least 7 hours of constitutive operation to identify each once. Of course identifying every peptide in a solution isn't required to learn about the proteins in the sample, but this illustrates that speed alone will not immediately solve the challenge. The other factors, like perfect separations and high identification rates, represent greater challenges to solve.

Another approach in solving the time-management challenge is to better allocate the mass spectrometer resources to identify the "most useful" parts of the samples. Effective allocation requires information, and our approach provides the mass spectrometer with more options and information through software modifications and real-time data analysis. Here, the mass spectrometer can gain information about the sample in an automatic and dynamic fashion, and can change course when it sees fit. Software improvements are ideal, since they cost nothing to deploy and can modify existing instruments without hardware upgrades. However, since intelligent methods are still in their infancy, much work needs to be done. The following sections outline the two largest hurdles preventing widespread use of IDA methods. First, there must be improvements made to how the analysis is conducted and how to respond to the results. Little has been done in this regard. IDA methods need to demonstrate substantial improvements over other techniques before more researchers will use them. The other challenge for IDA methods is that they lack general accessibility. It is difficult to implement such methods on your own, and instrument vendors have been reluctant to distribute them. Changes need to be made on how instrument vendors provide access to new methods before they see wider use.

Improving Decision Making in Intelligent Data Acquisition

Increasing the use of IDA requires proving and improving its usefulness to other researchers. If some other method can accomplish the same task, or do it better, then IDA methods will not be used. We have demonstrated that IDA methods can improve certain aspects of data collection. In chapter 2, several types of improvements made by IDA were outlined, such as improved quantitation for isobaric labels and SILAC, as well better PTM localization. Chapter 3 discussed increases in run-to-run reproducibility of peptide identification. We believe that IDA methods are capable of improving data quality and throughput in other areas as well. However, additional work needs to be done to 1) make IDA methods even better than traditional methods (e.g., DDA and DIA), 2) improve the algorithms used to analyze spectra and 3) make smarter real-time decisions.

To allocate resources efficiently and maximize data quality, the mass spectrometer needs the best available information in the shortest amount of time. This involves designing algorithms to analyze spectra quicker and more accurately. Any delay caused by data-analysis further hinders IDA methods compared to more traditional methods—which due to their simple construction take minimal time to execute. Unfortunately, the methods described in this document were developed using the ion-trap control language (ITCL) which lacks many features. One missing feature that greatly hinders IDA methods is the lack of asynchrony—only one thing can be done at a time. The MS could not set variables for the next scan while analyzing the previous spectrum, and it would have to wait till the analysis step was complete to start the next scan. This considerably slows down the instrument acquisition rate. Developing a system where the instrument duty-cycle is not negatively affected by the real-time data analysis is a very important step to improve the results.

Improvements must also be made to the decision making steps that follow real-time analysis. There is no benefit in analyzing a MS spectrum in real-time if there is no response to the results. Appropriate responsive action is necessary to improve data quality. Deciding what to do and when to do it becomes one of the biggest challenges to IDA methods. For example, in the middle of a LC-MS/MS experiment, *inSeq* identified a peptide with a post-transitional modification from a MS/MS spectra. However, the spectral quality was not good enough to localize the PTM. What should the mass spectrometer do next? Resample it with a different dissociation technique? Increase the resolution? Finding the answer to this and other possible scenarios is an important part of IDA and needs to be more fully explored. The work described in this document only briefly explored possible actions and a lot more work can be devoted to increasing this aspect. The responses also may be dependent on the sample, or the type of analysis being performed, and may change from experiment to experiment. So providing a robust set of options that can cover a multitude of experimental conditions is challenging.

Accessibility of Intelligent Mass Spectrometers

The other issue that faces intelligent data acquisition methods is the lack of general availability to researchers. Enabling new methods requires modification of the instrument control logic, which is not always straightforward to implement. There are two ways for increasing the intelligence of mass spectrometers. The first would be to modify a home built mass spectrometer, where the researcher has full access over the control logic. The other way is to modify and extend commercially available mass spectrometers with the desired abilities. For large-scale proteomic work, the former approach is not straightforward, as a vast majority of publications use commercial instruments for data acquisition. Custom built mass spectrometers often focus on a very specific task (e.g., mass analyzer development, new dissociation techniques, etc.) and are rarely geared for high-performance, large-scale LC-MS protein experiments. Even if a researcher built a mass spectrometer capable of these types of experiments, there is no easy way to disseminate the technology, short of starting a company themselves and selling their work. On the other hand, commercial instruments are primarily developed to take the best technologies available and combine them into one unified package. This results in a powerful and stable instrument that can handle the largest experiments. However, in order to protect their intellectual property (IP), instrument vendors are usually highly restrictive in how their instruments are used and modified. This makes implementing novel acquisition methods very difficult, and therefore general acceptance of these methods is slow. Thus, increasing the accessibility and availability of IDA methods is the a very important factor in its future use.

Probably the best way to propel the development of intelligent acquisition methods forward is to increase its accessibility and availability to researchers. This is challenging since instrument vendors are highly protective of their products; they have to protect their intellectual property and public imagine while providing state-of-the-art technologies to consumers. They are wary of providing access to their control logic for fear of competition. They also worry that supporting third-party programs for their instruments could damage their reputation if things go wrong. Our lab, which has developed multiple technologies now commercialized, knows first hand the care instrument vendors take in releasing third-party technologies to the general consumer. The following section will briefly discuss how new technologies are currently developed and suggests improvements to facilitate the dissemination of intelligent mass spectrometers.

Instrument Programming. To develop new MS instrument methods, researchers are typically given special access to the instrument's firmware by a vendor. This allows them direct control of the instrument and gives them the ability to alter the methods as they see fit. This is a burdensome process, as developing software in the firmware of a MS instrument is difficult to do and test. The programmer has to spend a good deal of time understanding the firmware code that controls the MS before development on new methods can begin, often without documentation. Also, firmware modifications is notoriously difficult to test and debug, especially on the LC-time scale. With no way to debug, samples and experiments must be conducted in full to test the change of a single variable; this is a very slow process of programming. If there are any bugs in the code, hours could be wasted trying to detect and locate them. Improvements in how new methods are made are needed to make developing a faster and more productive process.

Distributing the instrument's firmware is also not an ideal way of providing access and is more of temporary fix than a real solution. Most instrument vendors never developed a system to support third-party methods, so when the first researchers wanted to extend the instrument capabilities, the quickest solution was to give them access to the source code, just like they were an internal developer. After they have developed their technologies, in order for other research groups to use them, the developers had to work with the vendors to commercialize their products. This process is ineffective and slow, and prohibits mass distribution.

A better model of development would be to develop and deploy an application programming interface (API) to control the instrument. Here, the vendor would provide a set of software tools and objects to enable access to different parts of the instrument. They would have fine-grain support on what they make available to the end user and what they keep hidden, thus alleviating some IP concerns. This technique also could provide error checking, preventing the user from setting some value that could potentially damage the instrument or injure someone. In fact, some instrument vendors have started down this path. Thermo has recently released an API for their Q Exactive mass spectrometer to enable third-party support. This allows the user to program in an more advanced language than is used on the instrument itself. For example, the Q Exactive's firmware is written in Python (a scripting language), while the API is written in C# (a compiled, and generally a more powerful language). This also makes programming the instrument compatible with libraries such as CSMSL discussed in Chapter 4. An API model also allows users to share their code without IP issues, and could greatly improve the code when multiple developers are working together.

The biggest challenge is convincing the instrument vendors to support such a technology, as it requires time and resources to develop and maintain. But if anything can be learned from community-developed applications on the internet (i.e., crowd sourcing), much can be gained when many people are working on a common problem. It may well behoove vendors to provide such access to potentially gain dozens of developers. The other issue is the distribution of software to other researchers. An ideal solution would again mimic the internet, by constructing a central marketplace to download and install methods. This would greatly facilitate the distribution aspect and improve accessibility.

COLOPHON

This document was typesetted with LaTeX 2_εusing the MiKTeX project. It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).