Bioinformatic Solutions to Complex Problems in Mass Spectrometry Based Analysis of Biomolecules

Ryan M. Taylor
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Biochemistry Commons, and the Chemistry Commons

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/5585

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Bioinformatic Solutions to Complex Problems

in Mass Spectrometry Based

Analysis of Biomolecules

Ryan M. Taylor

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

John T. Prince, Chair
Mark Clement
Steven W. Graves
Barry M. Willardson
Dixon J. Woodbury

Department of Chemistry and Biochemistry
Brigham Young University

July 2014

Copyright © 2014 Ryan M. Taylor
All Rights Reserved
ABSTRACT

Bioinformatic Solutions to Complex Problems in Mass Spectrometry Based Analysis of Biomolecules

Ryan M. Taylor
Department of Chemistry and Biochemistry, BYU
Doctor of Philosophy

Biological research has benefitted greatly from the advent of omic methods. For many biomolecules, mass spectrometry (MS) methods are most widely employed due to the sensitivity which allows low quantities of sample and the speed which allows analysis of complex samples. Improvements in instrument and sample preparation techniques create opportunities for large scale experimentation. The complexity and volume of data produced by modern MS-omic instrumentation challenges biological interpretation, while the complexity of the instrumentation, sample noise, and complexity of data analysis present difficulties in maintaining and ensuring data quality, validity, and relevance. We present a corpus of tools which improves quality assurance capabilities of instruments, provides comparison abilities for evaluating data analysis tool performance, distills ideas pertinent in MS analysis into a consistent nomenclature, enhances all lipid analysis by automatic structural classification, implements a rigorous and chemically derived lipid fragmentation prediction tool, introduces custom structural analysis approaches and validation techniques, simplifies protein analysis form SDS-PAGE sample excisions, and implements a robust peak detection algorithm. These contributions provide improved identification of biomolecules, improved quantitation, and improve data quality and algorithm clarity to the MS-omic field.

Keywords: bioinformatics, quality assurance, mass spectrometry, lipidomics, proteomics, machine learning, lipid fragmentation, simulated dataset
ACKNOWLEDGEMENTS

To my parents for instilling a thirst for knowledge, my collaborators for sharing their knowledge, experience and judgement, my wife for all her support, my coworkers for reminding me why I enjoy research, my mentors for their guidance and counsel, particularly John who has always valued my growth first and foremost, both intellectually and spiritually, and my Heavenly Father who has never given up on me.
# TABLE OF CONTENTS

Bioinformatic solutions to complex problems in mass spectrometry based analysis of biomolecules ........................................................................................................................................... i

ABSTRACT .......................................................................................................................................... ii

ACKNOWLEDGEMENTS ................................................................................................................... iii

TABLE OF CONTENTS ..................................................................................................................... iv

LIST OF TABLES ..................................................................................................................................... viii

LIST OF FIGURES ................................................................................................................................... ix

Chapter 1  Introduction ................................................................................................................... 1

1.1 Biomolecules make life possible .......................................................................................... 1

1.2 The rise of ‘omics .................................................................................................................. 1

1.3 Mass Spectrometry enables analysis of biomolecules ....................................................... 2

1.4 State of Proteomics ............................................................................................................... 4

1.5 State of Lipidomics .............................................................................................................. 4

1.6 Data analysis is the major bottleneck ................................................................................... 5

Chapter 2  Improving nomenclature to uniquely map molecular entities to mass spectrometry signal ................................................................................................................................. 9

2.1 Abstract .................................................................................................................................. 9

2.2 Background .......................................................................................................................... 10

2.3 Results and discussion ......................................................................................................... 25
Chapter 3  Sequence and Structural Characterization of Great Salt Lake Bacteriophage CW02, a Member of the T7-Like Supergroup

3.1 Abstract

3.2 Introduction

3.3 Materials and methods

3.4 Results and discussion

Chapter 4  Resolving double disulfide bond patterns in SNAP25B using liquid chromatography–ion trap mass spectrometry

4.1 Abstract

4.2 Introduction

4.3 Materials and method

4.4 Results

4.5 Discussion and conclusions

4.6 Acknowledgements

Chapter 5  Automatic lipid classification by machine learning

5.1 Abstract

5.2 Introduction

5.3 Methods
Chapter 6  Msplinter: a structure driven approach to lipid fragmentation prediction............105
  6.1 Abstract ......................................................................................................................105
  6.2 Introduction .................................................................................................................106
  6.3 Methods .....................................................................................................................108
  6.4 Results .......................................................................................................................113
  6.5 Discussion ..................................................................................................................116
  6.6 Conclusions ...............................................................................................................119

Chapter 7  Massifquant: open-source Kalman filter based XC-MS isotope trace feature
detection ..........................................................................................................................120
  7.1 Abstract ......................................................................................................................120
  7.2 Introduction ................................................................................................................121
  7.3 Methods .....................................................................................................................124
  7.4 Results .......................................................................................................................133
  7.5 Discussion & Conclusions .........................................................................................139

Chapter 8  Metriculator: quality assessment for mass spectrometry-based proteomics .........145
  8.1 Abstract ......................................................................................................................145
  8.2 Introduction .................................................................................................................145
  8.3 Software ......................................................................................................................147
8.4 Ms-archiver—integration with a workflow ................................................................. 149

Chapter 9 Mspire-Simulator: LC-MS Shotgun Proteomic Simulator for Creating Realistic
Gold Standard Data ............................................................................................................ 150

9.1 Abstract ....................................................................................................................... 150

9.2 Introduction .................................................................................................................. 151

9.3 Methods ...................................................................................................................... 153

9.4 Results ....................................................................................................................... 154

9.5 Discussion and conclusion ....................................................................................... 163

Chapter 10 Conclusion .................................................................................................... 168

References ....................................................................................................................... 171
LIST OF TABLES

Table 3.1: CW02 putative ORFs and analysis of proteins determined by LC-MS and detected homologues. ........................................................................................................................................ 46

Table 4.1: The number of m/z fragments predicted for each fragmentation scenario in the SNAP25B cysteine-rich region........................................................................................................................................ 83

Table 5.1: Classifier performance for entire LMSD and categories slices of the LMSD. ..........101

Table 6.1: Comparison of approaches for MS/MS lipid fragmentation prediction algorithms. ..106

Table 6.2: Collision induced dissociation ions characteristic of selected Ceramides from Hsu et al., 2002. ............................................................................................................................................ 114

Table 6.3: Msplinter prediction scores of 7 experimental spectra against isobaric species +/- 5 amu from the experimental peak. .................................................................................................... 115

Table 7.1 centWave optimization on MM14 improved with identification performance and the parameters are in the same vicinity. ................................................................................................ 133

Table 9.1: Statistics comparing the two features shown in Figure 9.5 .......................................... 165
LIST OF FIGURES

Figure 1.1: Rise of the omic research fields..............................................................................................2

Figure 2.1 Ambiguity in current nomenclature. .......................................................................................16

Figure 2.2: The PSI CV states that a peak is “A localized region of relatively large ion signal in a mass spectrum.” ........................................................................................................................................17

Figure 2.3: The PSI controlled vocabulary has over 2,400 entries..........................................................18

Figure 2.4: The terms profile and centroid in combination with the other terms proposed allow distinction........................................................................................................................................20

Figure 2.5: In this partial overview of the proposed nomenclature, the relationship between base concepts and some qualifier terms is demonstrated. ........................................................................26

Figure 2.6: Integrated isotopic trace ........................................................................................................28

Figure 2.7: Deisotoped isotopic envelope ..................................................................................................30

Figure 2.8: Deisotoped molecular envelope .............................................................................................32

Figure 3.1: The conserved structure module arrangement of the T7 supergroup. ..........................56

Figure 3.2: Protein composition of mature wild-type CW02 particles ...............................................57

Figure 3.3: Coiled-coil prediction profiles of CW02 gp49 and the scaffold protein of phage T7 (gp9). ......................................................................................................................................60

Figure 3.4: Electron microscopy of CW02 by negative stain (A) and cryogenic (B) methods. ....62

Figure 3.5: Cryo-EM-based reconstruction of CW02. ...........................................................................63

Figure 3.6: Rigid-body fit of the HK97-like fold into the CW02 capsid. .................................................65

Figure 3.7: Segmented views of the subunits that make up the CW02 (A) and lambda (B) capsids. ......................................................................................................................................69

Figure 4.1: Disulfide bond patterns in SNAP25B. ......................................................................................73
Figure 4.2: CID cleavages observed for test peptides........................................................................79
Figure 4.3: MS/MS spectra of the oxidized peptide FCGLLVLPCK (C88L C90L mutant). ....80
Figure 4.4: MS/MS spectra of the oxidized peptide FLGLCVCLPK (C85L C92L mutant). ....82
Figure 4.5: Decision tree logic supporting the correct identification of each disulfide pattern. ...84
Figure 4.6: Identification of the three different disulfide patterns..................................................87
Figure 4.7: MS/MS spectra of the precursor mass m/z = 591.73 at retention time 68.03 min.....88
Figure 4.8: MS/MS spectra of the precursor mass m/z=591.73 at retention time 71.94 min. ......89
Figure 4.9: MS/MS spectra of the precursor mass m/z = 591.73 at retention time 77.9min. .......92
Figure 5.1: Lipid classification workflow schematic. ......................................................................99
Figure 5.2: Representative decision trees for LMSD classifications. .............................................100
Figure 6.1: An overview of the fragmentation schemes implementation. ........................................109
Figure 6.2: Comparison between CDF predicted fragment ions and published fragmentation
products (inset) for ceramide LMSP02010009. ..............................................................................116
Figure 7.1: Optimized performance metrics by dataset and algorithm. ..........................................134
Figure 7.2: A comparison of log-transformed percent quantitation errors (log ϵ) for successfully
identified ITs. ....................................................................................................................................135
Figure 7.3: A comprehensive view of manually annotated ITs on the MOUSE data set ............136
Figure 7.4: Isotope trace detection performance by quantiles for IT characteristics of the MOUSE
data set ...............................................................................................................................................138
Figure 7.5: Real-world application test. ..........................................................................................143
Figure 8.1: Visualization Plot example. .........................................................................................149
Figure 9.1: Overall process of simulation from theoretical spectrum to realistic peaks. ..........155
Figure 9.2: Elution peak shape. ......................................................................................................157
Figure 9.3: Intensity and m/z variance. .................................................................158

Figure 9.4: Visual output from the curve fitting algorithm. ........................................161

Figure 9.5: Comparison of simulated and actual MS features. ..................................162

Figure 9.6: Comparison between simulated and actual MS runs. ..............................163

Figure 9.7: Bird’s-eye view of a simulated complex human cell run. .........................164
Chapter 1 Introduction

1.1 Biomolecules make life possible

Nearly all known living organisms share certain common features, regardless of their diverse habitats. These common characteristics include: carbohydrates as storage molecules and signaling moieties; lipids as signaling molecules, structural components of membranes, and energy storage; nucleic acids as genetic storage and transmission molecules, as well as catalytic activity relating to gene expression; and proteins as molecular machinery which makes signal transduction, nucleic acid replication and transcription, and oxidative phosphorylation possible as well as forming structural components of both cells and tissue. These are the biomolecules which make life, as we know it, possible.

1.2 The rise of ‘omics

Biomolecules can be dauntingly diverse. While nucleic acids are composed from a relatively simple 4 letter alphabet of nucleotides from which all genetic material is composed, proteins are comprised of a ~20 letter alphabet of amino acids. However, due to the polymeric nature of macromolecules derived from amino acids and nucleotides, there is no limit to the possible molecular uniqueness. Carbohydrate derived macromolecules are polymeric as well, and due to the number of different ways they can be joined, they comprise an effectively limitless set of possible components and thus present with complexity orders of magnitude greater than that of proteins.¹ More than 35,000 lipids are structurally known, but the lipidome is estimated to contain more than 125,000 unique species. The metabolome refers to any cellular metabolite generally, and thus is comprised of chemically diverse small molecules which have highly variable physical properties, comprising some ~8000 species.²
Recent trends in scientific research demonstrate the drive to increase the rate of discovery. As grant acceptance rates have dropped by 50% in the last 8 years, scientific research can be facilitated by reducing costs, especially by using the greater data gathering capacity of omic style experimentation to complement more established procedures. Analysis of keyword frequency amongst published scientific research curated in Pubmed demonstrate the dramatic rise of omic publications in primary literature (see Figure 1.1).

Genomics has enabled the rise of modern proteomics. Original genomics projects were laborious undertakings until the development of high throughput instrumentation (next generation sequencers) which enable routine, organism level genomic analysis. Proteomics, lipidomics, and glycomics have required similar, high throughput instrumentation.

1.3 Mass Spectrometry enables analysis of biomolecules
In 2002, the Nobel prize in chemistry was awarded for the development of mass spectrometry (MS) ionization methods which are compatible with biomolecules. These innovations have enabled modern identification analysis of biological macromolecules.
Mass spectrometry analysis is typically performed in two steps: survey scans, and fragmentation scans. Survey scans determine all ions which are being introduced to the instrument at any given point. These scans are often acquired at high resolutions which ensures that all ions are separately identifiable in the spectra. Survey scans show the greatest number of species, as even intensity differences of several orders of magnitude can be detected. However, not all ions can be distinguished by survey scan alone. Some of these identification collisions arise by instrument limitations; that is without infinite resolution, some non-isomeric compounds cannot be differentiated. Isomeric amino acid residues (leucine and isoleucine) and other isomers are also indistinguishable. Even with the very best in instrumentation, at a mass accuracy of < 1 ppm in survey scans, mass accuracy is not sufficient to exclude alternative ion identities from subsequent analysis for more than 60% of possible peptides.\textsuperscript{5,6} To increase the coverage of analysis, orthogonal separation techniques are employed to reduce the complexity of ions eluting to the mass analyzer concurrently, including: liquid chromatography, 2D liquid chromatography, offline fractionation, and 2D gel filtration separations.

Fragmentation scans are performed by isolation of selected ions discovered in a survey scan for subsequent fragmentation. This often produces characteristic fragmentation patterns to aid in identification of biomolecules. These secondary (MS\textsuperscript{2}) scans can provide diagnostic ions which describe the structural characteristics of the ion. Some instrumentation is capable of subsequent rounds of fragmentation, MS\textsuperscript{N}, to further differentiate similar compounds and to produce additional diagnostic ions. Fragmentation scans can typically be acquired at a rate of several per second, although some instruments acquire them at 30+ scans/second. Fragmentation data is
capable of elucidating complex structural characteristics and can provide robust identifications for any type of biomolecule.

1.4 State of Proteomics
Proteomics is the most advanced MS based omic field. Proteomics is successful because 1) protein digestion into peptide fragments simplifies downstream analysis; 2) proteins are a product of the genetic code, which has provided, since 2001, a human protein dictionary; and 3) the basic rules which govern protein fragmentation are well characterized and understood, and fairly simple. Proteomics has enabled identification of 10,000 proteins from a single sample, and routinely provides the capacity to elucidate identifying, structural, and even functional details of protein biomolecules.

1.5 State of Lipidomics
Lipidomics is a field which is confronting several major challenges. With the rise of mass spectrometry based omics, the diverse roles of lipids in physiology and molecular biology has become more appreciated. While lipids were thought to comprise some several hundred species at most only a few decades ago, conservative estimates suggest >125,000 distinct lipid species exist. Further, while genomics, transcriptomics, and proteomics are based on a defined vocabulary which provides both known identities and a sequence driven ontology, lipids are composed of diverse structures which are in no way constrained by nomenclature or repeatable structural units. While most lipids originate from two biochemical pathways, the breadth and diversity of their structural features arise from the multitudinous enzymatic pathways devoted to their modification, cyclization, and oxidation. Lipidomic studies remain especially challenging as analysis tools are still in their infancy, and are not yet capable of even routine
These limitations are understandable given the expanded complexity relative to other biomolecules studied by MS analysis.

1.6 Data analysis is the major bottleneck
Despite the considerable advantages of MS based omic studies the primary bottleneck is the analysis of the massive volume of data generated. Proteomics has been benefitted by large scale, proprietary and open-source algorithms which provide capabilities of identification and even quantitation of labelled and label-free data. Even so, proteomics data analysis only identifies some 60% of MS/MS spectra, which are only ~10% of the ions observed in survey scans. The limitations of lipidomic and metabolomics analysis further demonstrate the limitations of current data analysis techniques in predicting complex ion fragmentation behavior within mass spectrometers.

Data analysis challenges in MS fields are primarily related to identification or quantitation. One challenge arises because of a lack of an established canonical terminology for data analysis algorithms. Even for the tools which are published, many are never utilized and comparisons are rarely published, leaving even superior algorithms relatively poorly utilized. Nomenclature used in MS have been standardized by IUPAC and HUPO-PSI, but these terms are largely designed for use in wet lab sample preparations and instrumentation descriptions. We introduce a MS algorithm nomenclature solution in Chapter 2.

1.6.1 Challenges in Identification Analysis
Traditional identification workflows depend upon known gene coding regions of a genome. The known protein sequences enable routine analysis of the protein from a complex biological sample. We present an example of this workflow in Chapter 3 in application to a novel virus
characterization. In this example, we also demonstrate the challenge of identifying which open reading frames discovered by analysis of the viral genome encode protein.

While protein identification has largely become routine, studying structure by MS remains quite challenging due to the large number of combinations possible, the relative reduction of ion counts, and increased overlap between fragmentation products. Interactomics, or the study of protein-protein interactions, often relies on crosslinkers to connect proteins which interact. The complex fragmentation patterns exhibited by these crosslinked proteins make identification nontrivial. Further applications lie in analysis of cysteine rich regions of proteins, which are often indicative of highly important structural characteristics. By careful optimization and iterative development coupled to customized data analysis workflows, complex and challenging structural details can be elucidated by mass spectrometry as demonstrated in Chapter 4. I was also involved in the characterization of a chaperonin complex by MS analysis which further demonstrates the ability of a targeted analysis to provide meaningful structural information.

Successful lipid analysis fragmentation prediction algorithms have relied upon ‘divide-and-conquer’ approaches to segment the lipidome and therefore reduce the required search space. Lipids are classified by a manually curated lipid classification system introduced by the LIPID MAPS consortium which provides a systematic and unified mechanism for discussing lipids specifically and unambiguously, and a reasonable system for segmentation of the lipid fragmentation problem.

In Chapter 5 I introduce an on-the-fly lipid classification system that empowers complete lipidome search tools to segment the fragmentation space even for novel lipids.
1.6.2 Challenges in Quantitative Analysis

Quantitation of MS datasets is necessary for comparative studies and is typically accomplished by either 1) label based quantitation, or 2) label-free quantitation. Label based methods rely upon introduction of covalent or isotopic modifications to the analytes directly which then enable differentiation of identified species between one or more experimental conditions. These labels provide reporter ions or mass shifts to analytes to determine which experiment they represent. Labeling can provide absolute quantitation, but decreases the rates of identification, greatly reduces ion intensity, and requires experimental conditions capable of isotopic label introduction or sample treatment labeling.

Label free quantitation methods require no special sample preparations or considerations experimentally and can therefore provide quantitation upon any sample. These methods are capable of both relative and absolute quantitation between case and control. While several competing implementations exist, the best performers consider peak identification and area under a curve as quantifying characteristics and thus enable robust and accurate quantitation with no special effort beyond data analysis. Many of these implementations are benefited from more accurate peak extraction algorithms. We introduce one implementation and comparison to existing algorithms in Chapter 7.

1.6.3 Challenges in Quality Control

The tremendous capabilities of modern mass spectrometers are entirely dependent upon the performance of the instrumentation. Assessing the quality of a produced dataset is very laborious and typically relies upon experienced technicians. Mass spectrometers are known to exhibit variable performance at times. Many instruments rely upon subjective evaluation of quality by instrument operator based on limited data sampling and inspection of visual indicators.
of quality. Hence, operator time is forfeit and scientific reproducibility are suspect under such constraints. In Chapter 8, I describe a complete automated workflow for quality control tracking and an encouragement to MS scientists to practice better performance monitoring.

Data analysis toolkit comparisons are necessary as means to select the highest quality analysis method and therefore ensure the greatest accuracy. However, the typical methodologies are dependent upon either human curation of annotated ‘gold standard’ datasets, or combining previous analyses to form a crowd-sourced truth. While these methods give some understanding of relative performance, a better mechanism is to simulate a dataset where the actual identities and quantities are known a priori as we introduce in Chapter 9.

Quantitative and reproducible metrics of instrument performance and simulated ‘known’ data sets for data analysis quantitation and comparison together provide capabilities to benchmark both instrument and data analysis performance to ensure conclusions drawn from data are valid.
Chapter 2  Improving nomenclature to uniquely map molecular entities to mass spectrometry signal

Author’s Note: This chapter outlines unambiguous terminology for MS based data analysis. I contributed to the ideas and preparation of the manuscript. This document is in submission to BMC Bioinformatics.

2.1 Abstract

Background: The comparison of analyte MS1 signal is central to many proteomic (and other -omic) workflows. Standard vocabularies for mass spectrometry exist and provide good coverage for most experimental concepts, however their terms for data concepts and algorithms are either ambiguous or missing. Without a standard, unambiguous nomenclature, literature searches, algorithm reproducibility, and algorithm evaluation for MS-omics data processing are nearly impossible.

Results: We show how terms from current official ontologies are too vague or ambiguous to explicitly map molecular entities to MS signals, and we illustrate the inconsistency and ambiguity of current colloquially used terms. We propose a set of terms for MS1 data processing which consists of a limited number of base terms along with qualifier terms allowing a vast number of MS1 data concepts to be succinctly, precisely, and intuitively described.

Conclusions: We suggest this nomenclature as a beginning to, not the culmination of, the standardization process.

Keywords: Controlled vocabularies; LC-MS; Peak detection; Feature detection; Proteomics

---

1 Publication authors are Smith R, Taylor RM, and Prince JT.
2.2 Background

Liquid-chromatography mass spectrometry (LC-MS) is a ubiquitous platform for proteomic (and other "omic") investigations. MS signal from hundreds to millions of ions can be quantitatively compared across experimental conditions in a fairly robust and repeatable way. Analyte quantities are captured directly in MS signal (aka MS1), while analyte identities are often elucidated or confirmed using MS/MS (aka MS2) fragmentation spectra. Confidently matching MS1 analyte signal between runs ("correspondence") is difficult with complex samples, so a variety of approaches to circumvent this problem have been explored. Multiple reaction monitoring (MRM) can be effective for monitoring a relatively small number of pre-selected analytes with a high degree of confidence, but it is unsuited to discovery-based experiments. MS/MS based approaches (e.g. iTRAQ and spectral counting) are also popular alternatives. However, due to low MS/MS capture rates (10-20%) and low database match rates (<60%), MS/MS driven approaches lack sensitivity compared to MS1-based approaches. And, although a data independent acquisition (DIA) approach may address some of the sensitivity deficiencies of MS/MS for identification, DIA does not of itself address difficulties in correspondence and quantitation. Hence, despite the availability of alternative approaches, the ability to match MS1 signal across experimental conditions is still highly desired.

Numerous efforts, large and small, have focused on using MS1 signal to compare analyte quantities. Ideally, solutions would focus on one of the several complex individual steps for data processing. However, most are released as end-to-end solutions (e.g., SuperHirn, MaxQuant, XCMS, and Skyline). This makes comparison to other competing algorithms virtually impossible, and is at least partially responsible for the lack of critical evaluations in the literature, since testing a subcomponent of a full software system requires re-implementing that
portion of the algorithmic pipeline. Our awareness of this problem has been accentuated as we have recently undertaken a survey in each of several of the modular subproblems of LC-MS quantification, including isotope trace detection, isotopic envelope detection, and correspondence. When one must distinguish the algorithmic details of several methods, or worse, implement them in code, one becomes painfully aware of the ambiguity in the terms currently used in MS data processing descriptions.

The lack of standard terminology has stagnated LC-MS data processing progress. Without consistent, clear terminology researchers have no handles for searching the literature, requiring onerous literature searches which fail to capture all relevant publications. Besides lack of access to possibly improved results, this leads to massive duplication of effort and few cross-tool evaluations since researchers are unaware of related efforts. A well defined vocabulary and problem domain also encourage and aid new-comers to the field—which currently poses a significant learning curve\(^\text{30}\)—improving solutions to difficult data processing challenges. It is also much easier to re-implement solutions when both the what and how of a process are clearly understood. Hence, an increase in term clarity has immediate impact on reproducibility—a requirement firmly enforced for sample preparation and wet-lab processing protocols but which is almost completely unenforced for data processing descriptions.\(^\text{32}\) What's more, it is virtually impossible to glean the algorithmic differences from a paper without a clear nomenclature so as to be aware of statistical biases imputed into the results of the algorithm. Without understanding the assumptions made in the algorithm, a practitioner is likely to overstate the significance of and confidence in experimental results.
What about using terms from existing HUPO-PSI\textsuperscript{77} and IUPAC\textsuperscript{78} controlled vocabularies (CVs)? Current CVs fail to adequately describe key data processing concepts precisely, precluding experimental reproducibly. The terms they do contain are ambiguous at best. There is a general lack of granularity that creates a gaping hole that ought to be addressed. The details of what terms the PSI and IUPAC committees choose to fill the present need will be a long process. As a small but critical first step towards eventual standardization, we have identified core data concepts and algorithms necessary for MS data processing that fulfill the general system we propose. We also propose a nomenclature to describe them which is constructed of a limited number of base terms along with qualifier terms. This combinatorial design allows a vast number of data concepts to be succinctly, precisely, and intuitively described. With precise terminology in hand, we are then able to show how terms from current official ontologies are too vague to explicitly map molecular entities to MS signals, and we illustrate the inconsistency and ambiguity of current colloquially used terms.

2.2.1 Colloquial Terms and Usage

Inconsistencies and ambiguities in colloquial terms has long been recognized as a problem. Consider, for instance, the usage of two of the most common labels for MS-omics data concepts. These lists are by no means exhaustive in references or instances.

The term *feature* is used for:

- An *isotopic envelope*.\textsuperscript{73,79–81}
- A *deisotoped integrated isotopic envelope*.\textsuperscript{82,83}
- An *integrated isotopic trace*.\textsuperscript{31}

The term *peak* is used for:
- A profile.\textsuperscript{74}
- A centroid.\textsuperscript{80,84,85}
- An isotopic trace.\textsuperscript{79,80,85,86}
- A deisotoped integrated isotopic envelope.\textsuperscript{31,87}
- An isotopic envelope.\textsuperscript{80}
- An integrated isotopic trace.\textsuperscript{80,83}
- An isotope.\textsuperscript{88}
- An instantaneous isotopic envelope.\textsuperscript{89}

It should be abundantly clear that these terms convey very little useful information—certainly insufficient information for reproducibility. Even terms with consistent use, there is a general lack of scope. For example, monoisotopic peak, meaning the most abundant isotopic trace in an isotopic envelope,\textsuperscript{77} cannot convey exactly what level of data processing has been used on the signal. Do the authors mean the isotopic trace of the most abundant isotopic trace in an isotopic envelope,\textsuperscript{79} the integration of that trace,\textsuperscript{89} or the summation of an entire isotopic envelope into one centroid?\textsuperscript{80,90} All of these uses fit the original definition, but none are specific enough to readily discern from just the term.

These examples briefly illustrate the ubiquity of overloading (using one term to mean more than one concept). Overloading treats a term as a variable, whose meaning must be defined in detail for the scope of each publication it appears in. An adequate definition takes significant thought, some descriptive text, and usually a descriptive image. There simply isn't ample space in each manuscript to define a custom set of terms for MS-omics data processing. This results in insufficient definitions for terms or no definitions at all.
For example, the terms *isotopic peaks* and *isotopic multiplets* do not convey a clear meaning and are undefined in the manuscript where they appear.\(^9\) It is unclear if a *peak/multiplet* is dealing with an *isotopic trace*, an *integrated isotopic envelope trace*, a *max isotopic envelope trace*, or an *instantaneous isotopic envelope trace*. The paper describes a decharging algorithm for *isotopic envelopes*, but depending on what definition you adopt for these terms, you will get a very different result.

As another example, consider a review paper that describes the algorithmic composition of several approaches to data processing problems.\(^8\) To allow for the use of mathematical algorithm descriptions, the author provides a key where symbols are defined for certain MS data constructs. These include symbols for *peak* area, number of *chromatograms*, *peak* maximum, *peak end*, *peaks* detected in a mass channel, raw height of *peak*, and *peaks* detected in a *chromatogram*. But what is a *peak*? What is a *chromatogram*? As seen from the citations in this paper, these terms are not universally defined, and the author does not define them. Subsequently, the algorithms in the paper are irreproducible unless the reader is able to correctly guess the definition of these terms intended by the author.

Reproducibility is, in fact, at the heart of the nomenclature problem. An algorithm description is rendered useless if the data structure terms used within it are ambiguous or undefined. In a modular approach to pipeline algorithm creation and testing,\(^7\) data processing methods prior to the pipeline module of interest have to be exactly describable with concise terms. In evaluating algorithms, knowing the exact format of the input data informs interpretation of the algorithm. If the data format is known, as well as the process used to transform, segment, and/or reduce the original, one can know what biases are intrinsic in the algorithm, as well as immediately suggest
improvements. For example, an algorithm that uses the whole isotopic envelope signal has more information to distinguish differences in a correspondence task than one that has reduced the whole signal to a single point representation. Do current CVs capture the required degree of descriptive specificity?

2.2.2 Why current CV terms are insufficient
IUPAC\textsuperscript{78} and HUPO-PSI\textsuperscript{77} are organizations that specialize in standardizing nomenclature. Their significant and useful controlled vocabularies address all aspects of MS experimentation. To date, most of the thousands of terms in these overlapping controlled vocabularies are focused on wet lab protocol and instrumentation. Although there are a few terms relevant to data processing, they seem to be a mix of canonized colloquial terms and terms that came directly out of a specific software platform rather than a consistent, coherent, and intuitive system.

2.2.2.1 Current CV MS data processing terms are ambiguous and inconsistent.
The HUPO-PSI-MS OBO has more MS data processing terms than IUPAC. Most are generic to the point of extreme ambiguity. For example, the term \textit{mass spectrum} refers to any segment of data with m/z and abundance axes: “a plot of the relative abundance of a beam or other collection of ions as a function of the mass-to-charge ratio (m/z).” This could refer to a host of different data segmentations, and seems to be a synonym for another term, \textit{profile spectrum}, defined as “A profile mass spectrum is created when data is recorded with ion current (counts per second) on one axis and mass/charge ratio on another axis” (see Figure 2.1).
An equally ambiguous complementary term is provided to refer to the time and abundance axes: *chromatogram*, defined as “the representation of detector response versus time.” This definition is not scoped at all with respect to the molecular entities whose signals are being measured. The term applies equally at any scope. In other words, any plotted entity that shows RT vs intensity qualifies as a chromatogram. Likewise, the term *total ion current chromatogram*, defined as the “chromatogram obtained by plotting the total ion current detected in each of a series of mass spectra recorded as a function of retention time” fails to imply any sort of scoping, and, worse, can correctly apply to any entity that qualifies as a chromatogram.

Figure 2.1 Ambiguity in current nomenclature. Any of the boxed signal in this figure can correctly be called a chromatogram, a mass spectrum, or a profile spectrum as defined by existing CVs. Additionally, the IUPAC terms isotope pattern and isotope cluster can refer to the signal in box 1, 2, or 3. It is clear that these terms pertain to too many distinct concepts to be clear. Note that existing nomenclatures have no way of distinguishing between these distinct data concepts, which can include parts of one or more combined signals. Box 4 represents a temporal snapshot of the spectrum. Arrow 5 refers to the process of generating an integrated isotope trace.
The term *peak* is defined in the PSI CV as “a localized region of relatively large ion signal in a mass spectrum.” As defined this term cannot discriminate between a host of distinct data concepts (see Figure 2.2). Among other associations, the term *peak* is used as a qualifier to describe a process officially named *peak picking*, when profile data is converted to centroid data. Thus, a peak can be any signal region (large or small) that consists of one or more centroids, which means any size subset of any data in any projection could be called a peak. The term has absolutely no specificity. The term *area peak picking* has a very unclear definition: “spectral peak processing conducted on the acquired data to convert profile data to centroided data. The area defined by all raw data points that belong to the peak is reported.” Intuitively, one would assume that *area peak picking* has to have a different meaning than regular *peak picking*, yet the distinction is not evident from the definitions, given the only difference is the addition of “The area defined by all raw points that belong to the peak is reported.” What is meant by “peak”?

![Figure 2.2: The PSI CV states that a peak is “A localized region of relatively large ion signal in a mass spectrum.” Note that each of the numbered data segmentations in this figure qualify as a peak according to the PSI CV. In the literature, each of the illustrated concepts can also be found referred to as a feature.](image-url)
Reported to what? This is one example of how some of the PSI CV terms seem more like unrelated combinations of terms from specific software user manuals than a curated collection of data concepts.

It is unclear what a centroid spectrum is. The definition states "processing of profile data to produce spectra that contains discrete peaks of zero width. Often used to reduce the size of dataset." However, a spectrum, as defined, can only have two dimensions: m/z and intensity.

![Diagram of PSI controlled vocabulary](image)

**Figure 2.3:** The PSI controlled vocabulary has over 2,400 entries. Only ten or so of these entries provide data processing concepts related to MS processing. Far from trying to compete with PSI or IUPAC, which are full-purpose controlled vocabularies, our proposal in this manuscript is to question the small subset of terms they provide for utility in describing MS processing. We propose a set of terms for MS1 data processing constructed of a limited number of base and qualifier terms that allow a vast number of MS1 data concepts to be succinctly, precisely, and intuitively described. The current ambiguous nomenclature yields situations where a given term refers to more than one concept and/or distinct terms refer to the same concept. An unambiguous nomenclature is composed of terms which refer to only one concept each, and no two terms refer to the same concept.

Since a spectrum cannot have an RT dimension, a centroid spectrum must be the same thing as a peak. It is assumed that instead they mean a peak picked profile signal summed through RT. It's all very confusing.

Finally, the inconsistency of the PSI CV is apparent. For example, the term feature is used in at least 25 definitions, but it is never defined. It is used to refer to at least a few different concepts, including the idea of a program parameter (MS:1000498, MS:1001760, MS:1002426), the isotopic envelope (MS:1001826, MS:1001827, MS:1002163), a PSI-CV mass spectrum (MS:1002166, MS:1002167, MS:1002168), and probably others (probably because it is unclear
what is being described in the other 16 definitions). Another term, mass trace, is similarly used in several definitions yet never defined. The implied use, like feature, overlaps the definitions of peak and chromatogram, making for considerable ambiguity (see Figure 2.3).

In addition to these generic terms, the PSI CV provides two specific data concepts: deisotoping and charge deconvolution. Deisotoping is referred to as “the removal of isotope peaks to represent the fragment ion as one data point and is commonly done to reduce complexity. It is done in conjunction with the charge state deconvolution.” The concept described is worthy of a definition, but the one provided can be improved upon. A fragment ion is not a data signal, but a molecular object. However, deisotoping is an operation on a data signal. Additionally, this term should not be specific to MSn fragment ions, but also applies to non-fragmented MS1 data, such as an MS1 isotopic envelope. Our nomenclature expands this term to include the logical wider use. Charge deconvolution is defined as “the determination of the mass of an ion based on the mass spectral peaks that represent multiple-charge ions.”

Deconvolution is already a widely used signal processing term (also used in MS processing92) for resolving two overlapping signals into their constituent parts (see Figure 2.4, top right). The PSI definition redefines an already widely used term to mean something other than what it means in all other contexts. What's more, the definition focuses on mass determination, not signal manipulation. It should be replaced.

2.2.2.2 Current CVs don't describe all necessary data concepts for MS data processing.

19
Crucial concepts for MS data processing are missing from the current CVs. Most of these relate to more specific concepts at higher granularity than is currently offered by the PSI and IUPAC CVs. A rigorous CV should allow for unique terms to describe signals from a peptide/lipid/metabolite down to individual data points. With the current nomenclatures, it is impossible to describe data processing algorithms' details using standard terms.

Explicit mapping of molecular entities to the signal their detected ions produce is essential to achieve clarity at all scopes of granularity. The PSI and IUPAC terms do not quite do this. For example, the best term to refer to a host of loosely related concepts is *ion*: “an atomic, molecular, or radical species with a non-zero net electric charge.” Ion is a proper and correct term that is

![Diagrams of isotope envelope profiles and centroids.](image)
general to science, and this is the widely used definition. However, this term applies to a charged item of any size, and cannot distinguish among the instances of interest (e.g. proteins in a proteomics experiment, lipids in a lipidomics experiment) and the smaller molecular charged units that are detected in a mass spectrometer experiment. How does one distinguish between the signal from the detection of a single isotope in a peptide, the signal from the detection of multiple isotopes of multiple instances of a given charged peptide, and the signals from all the charge variants of a peptide? What's more, as mentioned above, the term ion should never be applied to signal as it is in multiple instances in the PSI CV; an ion is a physical entity, and a signal can comprise a summary of the detection of one or more ions of one or more types (or noise). For instance, a standard isotopic envelope is the accumulation of signal detection from many instances of one type of molecule (say, a peptide) at one charge state.

2.2.2.3 *An incomplete CV impedes algorithm implementation/comparison.*

Writing code is a mathematically precise activity. Reproducibility requires the exact same equation, as it were, to be reproduced. Reproducibility requires specificity and clarity. Any ambiguity, overloading, or lack of detail makes the process impossible.

Frequently, it becomes necessary to code up a published algorithm. This could be because the algorithm was published independent of a software platform, or because you want to see the results of one particular module of a full-service program. As a case study, we recently attempted to code up one module of the MaxQuant algorithm in order to compare results of a feature detection algorithm. The following text is the pertinent portion, and demonstrates just how difficult it can be to parse through an algorithmic description without a good nomenclature, even in a well-written, top-tier-published manuscript:
...peaks are detected in a conventional two-dimensional (2D) way by first searching for local maxima of the intensity as a function of m/z. The lower and upper limits of the m/z interval for a 2D peak...are then determined by moving from the maximum position to smaller and larger m/z values, until either the intensity has dropped to zero, or a local intensity minimum has been reached....The centroid position of a 2D peak is then determined...If the peak consists of only one raw data point, then the m/z value of that point is taken as the centroid position. If there are two raw data points in a peak, then the centroid position is defined as the average of the two raw m/z values, weighted by the raw intensities....the 2D peaks in adjacent MS spectra are assembled into 3D peak hills over the m/z-retention time plane. Two peaks in neighboring scans are connected whenever their centroid m/z positions differ by less than 7 ppm. If for a given centroid in MS scan n no matching centroid is found in scan (n+1) in the ±7 ppm mass window, then it is checked if there is a centroid in scan (n+2) in the same mass window to continue the peak in time. We adjusted the window size to 7 ppm by visual inspection of many very low abundant peaks....A 3D peak is defined as the maximal chain of 2D peaks that results from connecting the centroids in time direction in the described way. At least two centroids have to be matched together to form a 3D peak, i.e. centroids that cannot be matched to centroids in the two previous or the two next scans are discarded....If a minimum is found whose value is 1/1.3 of the lower of the two local maxima the 3D peak is split into two at the minimum position...

With unclear terms, translating a manuscript into code is very difficult, and very unlikely to produce what the author intended.
Ambiguity also affects practitioners who do not write code; in the process of writing an exhaustive LC-MS correspondence survey, we recently parsed over 50 manuscripts describing correspondence algorithms, attempting to describe distinctions between them. Each manuscript had its own unique ill-defined vocabulary. It was very difficult to discern what data concepts the algorithm was operating on, which is essentially the bulk of the difference between those 50 or so methods. We had far more time for that exercise than a normal practitioner would when evaluating which method to employ in their experiment.

2.2.2.4 An incomplete CV impedes literature reviews.

Ambiguity also leads to manuscript bloat. While working on the correspondence survey, as well as several other surveys in progress, it became very apparent that most of the algorithms shared previously published approaches and failed to make a novel algorithmic contribution. This is a real problem, as it creates even more methods that must be compared against and even more papers to read. For example, there seems to be no difference between TracMass2—a recently published algorithm featuring isotope trace detection—and the uncited feature detection module of MaxQuant. Neither the authors nor the reviewers noticed the similarity of the two algorithms. As documented in this case and others described in, unintentional plagiarism is happening all over MS data processing. Clear nomenclature alleviates the burden of discerning differences in algorithms.

Having conducted extensive literature searches for LC-MS correspondence (published), isotope trace detection (unpublished), and isotopic envelope detection (unpublished), we can say with surety that the lack of unique terms makes the literature search nearly impossible to achieve a comprehensive search.
2.2.3 Community awareness and an immediate, intermediate response is the correct way of addressing this need. All controlled vocabularies are works in progress. Standards committees are best at crystallizing and refining accepted practice, but the onus to invent or select appropriate terminology lies foremost with the community itself. A good example of this was the creation of a standard spectrum exchange format. The mzXML format was created and published by a small group of researchers. After several years of use the HUPO PSI mass spectrometry working group produced the mzML format which was able to draw upon the experience gained from use of the mzXML format. Although a data format and not a CV, the success of mzML shows the good that can come of a manuscript-driven approach. mzXML was originally published as a manuscript, the sole product of a small group of researchers who noticed a problem and forwarded a solution. This was the genesis of community traction that culminated in the mzML standard, a significant step forward for all mass spectrometry users. An official nomenclature culminates with IUPAC and HUPO-PSI standards but the community cannot realistically expect nomenclature to begin there.

At present, the controlled vocabularies simply do not have coverage in the terms related to data processing. Because the problem is so extensive, and because opinions run strong in the domain of nomenclatures, this problem is best solved by drawing attention to shortcomings while providing a framework for unambiguous terms. A manuscript to draw attention to areas that can be improved are a viable means of correcting them, as demonstrated in other CVs. As we have shown by enumerating collisions, inconsistencies, and gaps in current terms, no small group of experts can successfully bring about a CV independent of an active, involved community, particularly when data processing represents only a small subset of the larger experimental
community. It is unfair to represent the data processing portion of the current PSI CV as the calculated and careful end product of a long, focused deliberation. Inspection of the terms makes it clear that this is not only a living, changing document, but (at least in terms of the data processing terms) seems to be, at least in part, an uncoordinated amalgamation of terms from different software groups. We submit that fostering community discussion in a peer-reviewed venue is at least as valid as open, uncoordinated, and seemingly minimally curated submissions to a standard.

2.3 Results and discussion
The PSI CV data processing terms are scant, inconsistent, and ambiguous for describing MS data processing. We propose a system for generating terms that allows greater unambiguous coverage of currently addressed concepts in a consistent and intuitive manner, facilitating reproducibility, comprehension, and searchability of data processing algorithms. The motivation at the heart of our proposed nomenclature is to explicitly map causal molecular entities to the signals they produce. In order to maximize the information communicated in a term, we have created base terms, which describe the general concept under consideration, as well as qualifier terms, which specify any additional information possible about the genesis of the data concept. An overview of all terms is presented in Figure 2.5.

2.3.1 Base Terms
Generic terms allow us to refer to a specific data structure without necessarily adding detail about how it was processed. These terms are useful for algorithms that will take data structured in a certain way, no matter how it came to be in the current format.
Figure 2.5: In this partial overview of the proposed nomenclature, the relationship between base concepts and some qualifier terms is demonstrated. The qualifier trace adds a time dimension to a base concept. An envelope is a set of related instances across the m/z dimension. An isotope is a molecule at a particular charge state with a certain number of neutrons. An isotopic envelope is the unique impulse signal (at a specific RT) generated by one molecule/charge state combination consisting of one or more isotopes equally spaced m/z 1/z apart. A molecular envelope is the set of unique isotopic envelopes generated by one molecule across multiple charge states. An isotopic trace is the unique whole (meaning throughout RT) signal generated by the accumulation of instances of a given molecule at a given charge state whose molecular formula contains the same isotopic composition. An isotopic envelope trace is the unique whole signal generated by one molecule/charge state combination consisting of one or more isotopic traces equally spaced m/z 1/z apart. A molecular envelope trace is the set of whole isotopic envelopes generated by one molecule across multiple charge states.
**Molecule** - The unit that accepts charge; a lipid in a lipidomics experiment or a peptide in a bottom-up proteomics experiment.

**Isotope** - An isotope in this context consists of a molecule, at a particular charge state, with a certain number of neutrons (no distinction is made in this context among molecules where the neutron is associated with different atoms or kinds of atoms) (see Figure 2.5).

### 2.3.2 Qualifiers

Obviously, the most specific term possible should be used in each instance. For this purpose, we have introduced a set of qualifying terms that add specificity to the above-defined generic terms. The use of qualifiers allows us to encode previous processing steps into the term used to identify a data structure.

**Profile**

Profile refers to the continuous signal produced by a mass spectrometer (see Figure 2.4). The qualifier profile allows us to specify concepts that are otherwise conflated between low-resolution and high-resolution data. For instance, in 2-d terms, an *isotope profile* is the data distribution thought to be a single isotope, and is found in high resolution profile data, while a *deisotoped isotope envelope* is the implicit convolution of several *isotope profiles* as a result of a low-resolution instrument (see Figure 2.4).

**Centroid**

Centroid is the qualifier that indicates a consolidation of the cumulative signal from an isotope profile into an impulse representation of the isotope, positioned at the center of mass of the profile signal. A convoluted profile signal, depending on how it is processed, may also produce multiple centroids (see Figure 2.4). This qualifier allows us to disambiguate between distinct
concepts such as an *isotopic envelope centroid* and an *isotopic envelope profile*. Note that the assumption is that all data is centroided unless otherwise stated and that the term only applies in the m/z dimension.

**Envelope**

An envelope connotes a discrete collection of things across the m/z dimension. For example, when we couple envelope with isotope, we get *isotopic envelope*, the unique impulse (meaning at a specific retention time (RT)) series generated by one molecule / charge state combination consisting of one or more isotopes equally spaced m/z 1/z apart (see Figure 2.5). By coupling *molecule* with *envelope*, we get *molecular envelope*, the set of unique isotopic envelopes generated by one molecule across multiple charge states (see Figure 2.5).
Trace

A trace implies a signal that extends into the RT dimension. For example, when we combine isotopic envelope and trace, we get an isotopic envelope trace, which is the unique whole (meaning throughout RT) accumulated (meaning throughout a run) signal generated by one molecule / charge state combination consisting of one or more isotopic traces equally spaced m/z 1/z apart (see Figure 2.5). Likewise, in an isotopic trace, the unique whole (meaning throughout RT) signal generated by the accumulation of instances of a given molecule at a given charge state whose molecular formula contains the same isotopic composition (see Figure 2.5). A molecular envelope trace is the set of whole (meaning throughout RT) isotopic envelopes generated by one molecule across multiple charge states (see Figure 2.5).

Integrated

An integrated object has been summed through the RT dimension. For example, if we take an isotopic trace and sum its constituent centroids (or profile points), we will end up with a single 3-tuple consisting of m/z, RT, and intensity that can accurately be called an isotope (see Figure 2.5). However, by calling it an integrated isotopic trace, we retain a unique description of the original data structure as well as the transforming process used (see Figure 2.6). An integrated isotopic envelope trace is the sum of the constituent points in the isotopic traces contained in the isotopic envelope trace. In appearance, it is identical to the isotopic envelope in Figure 2.5.
**Average**

The data concepts described by the qualifier average are, in appearance, the same as those in integrated, however the process to generate them involves taking the average of the intensity of the composite points, not the sum.

**Instantaneous**

The qualifier instantaneous implies that this object is a spectral slice of a trace object at a given RT. The instantaneous objects look exactly like those that are integrated; however, this qualifier indicates that we are looking at a slice of the data structure in time, not a summation or average of the data through time.

![Diagram](image)

*Figure 2.7: Deisotoped isotopic envelope - the composition of all isotopes in an isotopic envelope.*

**Max**

The qualifier max implies that this object is the spectral slice of a trace object at the RT of greatest intensity. Max objects look exactly like those that are integrated; however, this qualifier indicates that we are looking at a slice of the data structure in time, not a summation or average of the data through time.
Deisotoped

The qualifier *deisotoped* implies that the isotope envelope has been combined through the m/z dimension, such as a *deisotoped isotopic envelope*, the consolidation of all *isotopes* in an *isotopic envelope* (see Figures 2.7 and 2.8).

Reduced

The qualifier *reduced* implies that the object has been combined through reducing charge states to the lowest common charge state. For instance, a *reduced molecular envelope* is the set of the composition of all *isotopic envelopes* in the molecular envelope (see Figure 2.5, bottom left to middle left).

2.3.3 Clearer than colloquial terms

Our suggested vocabulary eliminates most if not all of the ambiguity in the current naming schemes employed in the literature. The following examples illustrate how the proposed vocabulary untangles the currently obfuscated terms in use.

*Isotopic envelope trace* describes a concept for which the following terms have all been used: an *eluting isotopic distribution*,\(^9\) a *chromatogram*,\(^8\) an *isotope series*,\(^8\) an *isotope pattern*,\(^7\) an *isotope-resolved mass spectrum*,\(^9\) an *ion series*,\(^9\) and an *isotopic cluster*.\(^9,9\) None of these terms differentiate between the concepts we refer to as *isotopic envelope trace, instantaneous isotopic envelope, max isotopic envelope*, etc.
Isotopic traces have been referred to as *eluting isotopes*, *single ion chromatograms*, *peaks*, *mass spectra*, and *peak hills*. Each of these terms are unclear. The problem with the term *chromatogram* is that it does not specifically refer to the elution profile of a single *isotope*. For example, an *extracted ion chromatogram* is an m/z slice of data that can extend across an entire run's RT. Any term that uses *peak* is bound to be confusing due to the overuse of the term. Like *chromatogram*, a *mass spectrum* can technically stretch across an entire m/z range and therefore does not specifically describe the m/z window of a specific *molecule*.

*Integrated isotopic envelope* has been called an *isotope pattern*. However, many other concepts can accurately be called *isotope patterns*, such as a *max isotopic envelope* or an *averaged isotopic envelope*.
Using this nomenclature, it is much simpler to clearly and unambiguously describe an MS data processing algorithm. The following text is the translation of the above quoted MaxQuant text translated into the proposed nomenclature.

Overlapping and/or contiguous profiles are deconvolved by bisecting all contiguous profiles with a local maximum bordered by local minima. Each profile is centroided by taking the weighted average of the m/z values of the points comprising the profile. Isotope traces are constructed from centroids by the following method: For each scan, each centroid within 7 ppm of an isotope trace from the previous scan or penultimate scan is aggregated to the closest (in ppm) isotope trace. All other centroids are considered new isotope traces for future scans. Any isotope traces with only one centroid after all scans are included are culled. A postprocessing mechanism to address erroneously appended isotope traces splits the trace anywhere a centroid is found with intensity less than or equal to 1/1.3 of the lesser intensity of two surrounding local maxima.

Not only does this text more readily reduce to code, it is easier to understand and takes up about half the text of the original. No term is used to mean more than one specific concept. The terms have a one to one mapping to the concept they refer to.

2.4 Expected Objections

Having discussed this nomenclature with many of our colleagues, we anticipate some objections and will address the most common here.

*This nomenclature competes with PSI/IUPAC.* We are not advocating for a replacement of either IUPAC or PSI controlled vocabularies, rather arguing that the subset of terms relevant to data
processing have insufficient coverage and are ambiguous where defined (see Figure 2.3). We are arguing that those terms relevant to data processing discussed here ought to be replaced, and those missing ought to be added.

*Why don't you submit these to PSI?* We are planning on doing so. However, this manuscript is not an attempt to change PSI. Standardizations do not drive the community, the community drives standardizations. There is a current dearth of appropriate terms to describe MS data processing. We have provided solid evidence that this is a problem, and we have proposed a nomenclature that solves that problem. We do not have a stake in PSI and cannot control their nomenclature. However, as a manuscript, individuals can begin using this nomenclature now no matter how long PSI takes to modify their CV (or whether they do or not).

*You have no data format.* The community does not recognize a controlled vocabulary and a data format (XML schema) as the same thing, as demonstrated by the fact that PSI has data formats and a CV, and they are separate products. The first step is establishing terms that describe these concepts. Producing a data format that represents these data concepts is a different problem that will have to be addressed in the future. Having a data format before establishing that this is a problem, let alone before coalescing on an industry-wide acceptable solution, is a mistake. It would require that any software tools coded between now and then be redone. Instead, we focus on the first problem, which is establishing that there is a problem. We propose a nomenclature, but expect and look forward to many constructive criticisms to improve upon it. Meanwhile, this manuscript serves as a cite-able lexicon for anyone who has the need to describe these concepts (for example, in an algorithm manuscript) yet has no available way of doing so within page limits.
You should be improving mzQuantML or mzIdentML instead of doing this. mzQuantML and mzIdentML are not CVs, they are data formats (see previous objection).

2.5 Conclusion
The ever-increasing number of MS-omics experiments drives a thriving MS-omics data processing algorithms field. However, the lack of an unambiguous vocabulary for MS-omics data concepts has created serious challenges for reproducibility and evaluation of data processing algorithms. In this paper, we have highlighted the ambiguity of current vocabulary for MS-omics data processing. We propose an unambiguous vocabulary together with a visual lexicon for the proposed terms. By adopting these terms, authors can facilitate reproduction of their algorithms succinctly by providing a crystal-clear set of meanings for terms they use, vastly improving the reproducibility of their work.
Chapter 3  Sequence and Structural Characterization of Great Salt Lake Bacteriophage CW02, a Member of the T7-Like Supergroup

Author’s Note: This chapter describes characterization of a novel halophage virus discovered in the Great Salt Lake. I contributed by development of the mass spectrometry sample preparation technique for gel fragment preparation for LC/MS, MS analysis, and in determining if the viral capsid contained anything other than DNA. These results were published in the Journal of Virology.

3.1 Abstract

Halophage CW02 infects a Salinivibrio costicola-like bacterium, SA50, isolated from the Great Salt Lake. Following isolation, cultivation, and purification, CW02 was characterized by DNA sequencing, mass spectrometry, and electron microscopy. A conserved module of structural genes places CW02 in the T7 supergroup, members of which are found in diverse aquatic environments, including marine and freshwater ecosystems. CW02 has morphological similarities to viruses of the Podoviridae family. The structure of CW02, solved by cryogenic electron microscopy and three-dimensional reconstruction, enabled the fitting of a portion of the bacteriophage HK97 capsid protein into CW02 capsid density, thereby providing additional evidence that capsid proteins of tailed double-stranded DNA phages have a conserved fold. The CW02 capsid consists of bacteriophage lambda gpD-like densities that likely contribute to particle stability. Turret-like densities were found on icosahedral vertices and may represent a unique adaptation similar to what has been seen in other extremophilic viruses that infect archaea, such as Sulfolobus turreted icosahedral virus and halophage SH1.

________________________

2 Publication authors are Shen PS, Domek MJ, Sanz-Garcia E, Makaju A, Taylor RM, Hoggan R, Culumber MD, Oberg CJ, Breakwell DP, Prince JT, and Belnap DM.
3.2 Introduction

Tailed double-stranded DNA (dsDNA) bacteriophages (order Caudovirales) account for nearly 95% of all documented phages and probably comprise the majority of the viruses on the planet. Recent estimations place the global number of virus particles on the order of $10^{31}$ to $10^{32}$, outnumbering the total number of host cells by a factor of 10. The ubiquity of phages suggests an ecological, predatory role in recycling organic matter, especially in extreme environments, where eukaryotic predators, primarily protozoal grazers, are rare. Phages also affect the genetic diversity of their environments by facilitating horizontal gene transfer between host cells. The sheer abundance of phages in aquatic environments may supply the greatest selective pressure exerted on host organisms.

Bacteriophages of the order Caudovirales are divided into three morphologically distinct families. All consist of isometric or prolate heads that encapsulate linear double-stranded (dsDNA) genomes, but phages of the Myoviridae family have a long, contractile tail (e.g., bacteriophage T4); viruses of the Siphoviridae family consist of a long, noncontractile tail (e.g., bacteriophage lambda); and phages within the Podoviridae family exhibit a short, noncontractile tail (e.g., bacteriophage T7). Recent high-resolution capsid structures of tailed dsDNA phages revealed a conserved fold that is shared among all three phage families. The first of these structures was solved for bacteriophage HK97, and subsequent structures of Caudovirales phages have been shown to adopt the HK97-like fold. The limited identity between primary structures of HK97-like capsid proteins is a prototypical example that structure is more conserved than sequence.
Halophage CW02 is a bacteriophage isolated from the south arm of the Great Salt Lake (GSL) and infects a euryhalophilic *Salinivibrio costicola*-like bacterium, SA50, also isolated from the GSL. At least two other tailed phages of *S. costicola*, UTAK\textsuperscript{109} and G3,\textsuperscript{110} have been reported; however, only general features of the structures and genomics of these phage are known. Here we characterized halophage CW02 to determine its relatedness to other dsDNA bacteriophages. Based on negative stain transmission electron microscopy, CW02 is morphologically similar to viruses of the *Podoviridae* family. Genomic analysis and mass spectrometry (MS) of CW02 genes and proteins revealed the strongest sequence similarities to *Pseudomonas* phage PA11,\textsuperscript{111} roseophage SIO1,\textsuperscript{112} vibriophage VpV262,\textsuperscript{113} and cyanophage Pf-WMP3,\textsuperscript{114} all of which are distant relatives of bacteriophage T7 but members of the T7 supergroup. The 16-Å resolution structure of CW02, solved by cryogenic electron microscopy (cryo-EM) and three-dimensional (3D) reconstruction, permitted fitting of the HK97-like fold into the capsid density, thereby supporting the observation that the capsid proteins of tailed dsDNA phages adopt a common ancestral fold.

### 3.3 Materials and methods

Cultivation and purification. Host bacteria and bacteriophage were isolated from the GSL in Utah. Sediment and water samples were collected between May and November 2008 in shallow waters along the north shore of Bridger Bay on Antelope Island of the GSL. Salinity was measured at 8% by evaporation. (Salinity is known to vary seasonally due to fresh water inflow and evaporation; see http://ut.water.usgs.gov/greatsaltlake/salinity/.) The pH ranged between 7.5 and 8. The host bacterium, SA50, was first isolated on Halobacteria Medium (per liter of solution, 80 g NaCl, 10 g MgSO\textsubscript{4} · 7H2O, 5 g casein hydrolysate, 5 g KCl, 3 g disodium citrate,
1 g KNO3, 1 g yeast extract, and 0.2 g CaCl2 · 6H2O), which was modified from the original 22% NaCl to 8% NaCl and pH 7.5.

The 16S rRNA gene of SA50 was amplified using bacterium-specific primers (27F, 5= AGA GTT TGA TCM TGG CTC AG 3=; 1492R, 5= ACG GYT ACC TTG TTA CGA CTT 3=). The reaction mixture contained 200 nM each primer, the deoxynucleoside triphosphates at 250 μM each, 0.2 mg ml⁻¹ bovine serum albumin, 1 U Taq DNA polymerase, and the diluted reaction buffer (Promega, Madison, WI). The amplification parameters were 94°C for 3 min; 25 cycles of 94°C for 45 s, 57°C for 1 min, and 72°C for 2 min; and a final extension step of 72°C for 7 min. Approximately 1,350 bp were sequenced (Molecular Research Core Facility, Idaho State University, Pocatello, ID) and queried against the GenBank database by use of the BLAST search tool. The isolate showed 99% sequence identity with S. costicola subsp. costicola strain ATCC 33508 (NR_027590.1). The isolate was maintained in Halobacteria medium.

CW02 was initially isolated by the use of a plaque assay of filtered GSL water with SA50 as the host strain. CW02 was recovered from plaques by removal from the soft agar overlay with a sterile Pasteur pipette. The agar was transferred to a sterile 8% salt solution, vortexed, and incubated for 30 min to release phage from the matrix. CW02 was then amplified by inoculation into a broth culture of SA50 and incubated on a shaker at 30°C. Phage replication caused lysis of the host cells and clearing of the broth culture after 18 h. After clearing, the broth culture was centrifuged at 5,000 × g for 10 min and passed through a 0.2 μm filter to remove bacteria. Host bacteria and bacteriophage filtrate stocks were stored frozen (~20°C) in 20% glycerol.

The bacteriophage was further purified by isopycnic ultracentrifugation (64% [wt/vol] CsCl) at ~125,000 × g for 24 h at 4°C. A single viral band was extracted from the gradient. CsCl was
removed by dialysis in either pure water or pH 8 buffer containing 1.35M NaCl, 48 mM MgSO$_4$·7H$_2$O, 1 mM CaCl$_2$, and 2 mM Tris-Cl. Finally, purified CW02 was washed and concentrated by centrifugal filtration (100-kDa molecular mass cutoff filters; Sartorius Stedim Biotech, Bohemia, NY) to be made suitable for cryo-EM.

3.3.1 Genome isolation and sequencing
The DNA was isolated as follows from 100 ml of bacteriophage filtrate. The filtrate was concentrated by centrifugal filtration (Amicon Centricon 100,000 molecular weight cutoff filter; Millipore Corp., Billerica, MA) to approximately 2 ml. Contaminant nucleic acids in the concentrated bacteriophage solution were removed by addition of DNase I (1 μg/ml) and RNase A (1 μg/ml) (Sigma-Aldrich Corp., St. Louis, MO), incubated at 37°C for 1 h, and then centrifuged at 110,000×g for 2 h. The pellet was suspended in 2 ml of SM buffer (5.8 g NaCl, 2 g MgSO$_4$·7H$_2$O, 50 ml 1 M Tris [pH 7.5], and 5 ml 2% gelatin per liter of distilled water) and treated with proteinase K (50 μg/ml; Sigma-Aldrich Corp., St. Louis, MO) and 0.5% sodium dodecyl sulfate (SDS; Sigma-Aldrich Corp., St. Louis, MO) at 56°C for 1 h to remove the capsid. The DNA was then extracted through phenol-Tris-Cl, followed by a 50/50 mixture of phenol-chloroform, and finally through pure chloroform. The DNA was ethanol precipitated and resuspended in Tris-EDTA buffer. Genomic DNA was sequenced at the Brigham Young University (BYU) DNA Sequencing Center using a Roche Genome Sequencer FLX instrument and employing the GS FLX Titanium Sequencing XLR70 kit (Roche Diagnostics Corporation, Indianapolis, IN).

3.3.2 Sequence analysis
Putative open reading frames (ORFs) of the CW02 genome were determined using GeneMarkS$^{117}$ and numbered according to whole-genome homology with bacteriophage
PA11. CW02 protein homologues were determined by PSI-BLAST analysis (http://blast.ncbi.nlm.nih.gov/Blast.cgi). All protein sequences for pairwise alignments were obtained from GenBank (http://www.ncbi.nlm.nih.gov/GenBank/), including sequences of enterobacterial phage T7 (NC_001604), VpV262 (NC_003907), SIO1 (NC_002519), Pf-WMP3 (NC_009551), and PA11 (NC_007808). Sequence alignments were performed using ClustalW2. Protein secondary structures were predicted using Psi-pred, COILS, BetaWrapPro, or Phyre.

3.3.3 Electron microscopy and image reconstruction
For negative staining, 3.5 μl of purified CW02 was adsorbed onto a glow-discharged, continuous- carbon-coated copper grid, blotted, and then washed. The grid was blotted again, stained with a solution of 1% uranyl acetate, and then blotted a final time and allowed to dry.

For cryo-EM, 3.5 μl of purified CW02 was placed on a glow-discharged, holey-carbon-coated copper grid, blotted, and plunge frozen in liquid ethane with an FEI Vitrobot (FEI, Hillsboro, OR). Specimens were transferred to a Gatan 626 cryoholder (Gatan, Inc., Pleasanton, CA) cooled with liquid nitrogen. Cryo-EM images were acquired at 200 keV and at ×39,000 magnification via low-dose methods at objective lens underfocus levels ranging between 0.2 and 2.8 μm. Focal pairs were taken 1.0 μm apart. Images were recorded either by a Gatan charge-coupled device camera (1,024 by 1,024 pixels) or on Kodak SO-163 film in an FEI Tecnai F30 transmission electron microscope (FEI, Hillsboro, OR). Electron micrographs recorded on film were digitized on a Nikon Super Coolscan 9000 ED scanner.

Particle images were extracted from scanned electron micrographs using X3DPREPROCESS. Contrast transfer function (CTF) signal and decay parameters were determined from and applied to the images by the use of Bsoft (bshow and bctf functions) and the algorithm of Conway and
Steven, except that images within a focal pair were not combined during CTF correction. Origins and orientations of the extracted particles were determined by use of the model-based technique of PFT2, which was adapted to use phase and amplitude information in orientation selection. A reference model for PFT2 analysis was generated by the random-model method with imposed icosahedral symmetry. The final 3D reconstructions were calculated using EM3DR2. In displayed images, contour levels are given in terms of $\sigma$, which was calculated as the number of standard deviations relative to the average map density.

For size calibration, poliovirus (160S form) was used as an internal standard. Poliovirus particles were mixed with CW02. The combined sample was imaged via cryo-EM, and separated images of poliovirus and CW02 were used to compute 3D image reconstructions. A previously calibrated map of 160S poliovirus was used to determine the size of the poliovirus reconstruction and, hence, the sampling size of the CW02 reconstruction. Bsoft (bshow function) was used to measure capsid dimensions.

3.3.4 Handedness determination
Tilt experiments were performed as described previously. Briefly, pairs of micrographs of CW02 were recorded with the same field of view, under the same conditions, but with the specimen at different tilt angles. The first micrographs of the tilt pairs were recorded at 0° tilt, and the second micrographs were recorded at 5° tilt. The origins and orientations of the untilted-particle images were determined using the PFT2 model-based method described above, after which the Bsoft program dhand was used to predict the orientations of the tilted particles (for either handedness), compute tilted projections from enantiomers of the 3D reconstruction,
and compare the projections to the experimental images of tilted particles. The correct handedness was the handedness that compared the best.

3.3.5 Structural modeling
The atomic coordinates of HK97 capsid protein\textsuperscript{134} (Protein Data Bank accession no. 1OHG) were fitted as separate pentameric or hexameric units. Only the C$\alpha$ chain of axial (A) and peripheral (P) domains were used in the fitting. Pentamers and hexamers were first fitted manually as a rigid body into the electron density map of CW02 by use of the program UCSF Chimera\textsuperscript{135} after which the fitting was refined by automated fitting using the same program. All surface renderings of the reconstruction were done at a 1-$\sigma$ contour level, which is defined by the sum of the average and standard deviation of map densities. Segmentation of individual subunits (via UCSF Chimera\textsuperscript{135}) was also carried out.

Attempts were made to reconstruct the turrets from cryo-EM images. We used a procedure adapted from a method developed previously by Briggs et al..\textsuperscript{136} After icosahedral reconstructions were computed, turrets were re-extracted from the 2D images using only vertices that did not overlap the densities of the capsid shell. Reference-free class averages (SPIDER\textsuperscript{137}), and asymmetric and symmetric (C2, C3, C4, C5, C6, C7; Cn is cyclic n-fold symmetry) reconstructions using random models as the initial reference, were computed. Additional class averages were also generated from cryo-EM pictures by masking the capsid shell according to the radius. In addition, we tested if turrets were found on all 5-fold vertices by performing asymmetric and C5-symmetric reconstructions using as the initial reference a CW02 capsid with a cylindrical density on a single vertex.
3.3.6 MS

CW02 virions were subject to SDS-polyacrylamide gel electrophoresis (SDS-PAGE) to separate individual proteins. The gel was stained using Coomassie blue reagent, after which individual gel bands were excised and then shredded by spinning gel fragments through micropipette tips. To capture all of the proteins in solution, whole virions were also prepared for MS analysis by forgoing SDS-PAGE. Proteins were prepared for analysis by MS by use of a modified filter-aided sample preparation protocol adapted from reference 57. Briefly, the gel fragments were destained using a 1:1 mixture of acetonitrile and 8 M urea, in 0.1 M Tris-HCl (pH 8.5). The denatured proteins were subsequently reduced with 0.1 M dithiothreitol in UA buffer (8 M urea, 0.1 M Tris-Cl, pH 8.5) and then carboxamidomethylated with 50 mM iodoacetamide in UA buffer. Each solution was washed by filtration through a 10-kDa molecular mass cutoff filter (Sartorius Stedim Biotech, Bohemia, NY). The urea and iodoacetamide were then replaced with 50 mM ammonium bicarbonate. Proteins were finally digested with sequencing grade trypsin (Promega) and then acidified by the addition of formic acid to 1% of the volume.

Acidified peptide samples were loaded onto a nanoAquity C18 column (75 μm by 15 cm; Waters Corporation, Milford, MA) and eluted by a 100-min binary gradient of Optima grade solvent A (5% acetonitrile, 0.1% formic acid) and Optima grade solvent B (0.1% formic acid, 99.9% acetonitrile) (Thermo Scientific, San Jose, CA) at 425 nl/min composed of the following steps. From a baseline of 95% solvent A, an 8-min gradient to 90% solvent A was performed, followed by a 65-min gradient to 65% solvent A, a 7-min gradient to 5% solvent A for 8 min, and a 3-min return to a 95% solvent A baseline for 9 min.
Column effluent was directed to a nanoelectrospray ionization source on an LTQ-Orbitrap XL (Thermo Scientific, Waltham, MA). Data-dependent acquisition was performed by coupling a 60,000 resolution survey scan in the Orbitrap with up to the top seven subsequent collision induced dissociation tandem MS (MS/MS) scans acquired in the ion trap (signal threshold of 1,000, normalized collision energy of 35%, isolation width of 2 \( m/z \), with a two-count dynamic exclusion in a 3-min window). A lock mass of a polysiloxane compound \( m/z \) 445.120025 was used as an internal mass calibration standard.\(^{138}\)

MS/MS spectra were converted to the Mascot generic format by the MsConvert program (http://proteowizard.sourceforge.net/pubs.shtml). Mascot searching was queried against a database of ORFs derived from the CW02 genome. The peptide search space was expanded by concatenating the CW02 database with a comprehensive database of the Escherichia coli ATCC 8739 proteome (GenBank accession no. YP_001723016.1). Search results were filtered against a decoy database to a 1% false discovery rate confidence at the peptide level, corresponding to Mascot ion cutoff scores of 22 to 27.

3.3.7 Nucleotide sequence and structural accession numbers
The CW02 genome, 3D reconstruction, and fit of HK97 coordinates into the CW02 capsid have been deposited in the GenBank database (JQ446452), the EM Data Bank (5388), and the Protein Data Bank (3J1A), respectively.
Table 3.1: CW02 putative ORFs and analysis of proteins determined by LC-MS and detected homologues. \(^a\)The first value in each column corresponds to data from SDS-PAGE band excision (see Figure 3.2). The second value, if present, corresponds to data from whole, non-SDS-PAGE-treated virions. \(^b\)Percentage of amino acid sequence observed. \(^c\)Summary of hits found via PSI-BLAST analysis. Protein sources are in parentheses. Conserved domains are underlined. \(^d\)Secondary structure prediction programs are underlined.

**Identification by LC-MS\(^a\)**

<table>
<thead>
<tr>
<th>ORF</th>
<th>Direction</th>
<th># amino acid residues</th>
<th>Molecular mass (kDa)</th>
<th>Unique peptides</th>
<th>% Coverage(^b)</th>
<th>BLAST hits and conserved domains(^c)</th>
<th>E value</th>
<th>Putative function and structure prediction(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp1</td>
<td>+</td>
<td>78</td>
<td>8.7</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF001 (<em>Pseudomonas</em> phage PA11) 7e-45 Amidotransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Glutamine amidotransferase (<em>Bacillus</em> sp. strain B14905) 1e-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>215</td>
<td>24.2</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF004 (<em>Pseudomonas</em> phage PA11) 4e-45 Amidotransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Glutamine amidotransferase (enterobacterial phage φEco32) 4e-21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>+</td>
<td>437</td>
<td>48.8</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF005 (<em>Pseudomonas</em> phage PA11) 1e-85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>+</td>
<td>64</td>
<td>7.2</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF005 (<em>Pseudomonas</em> phage PA11) 1e-85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>+</td>
<td>267</td>
<td>30.0</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF006 (<em>Pseudomonas</em> phage PA11) 6e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td>+</td>
<td>77</td>
<td>8.7</td>
<td></td>
<td></td>
<td>Conserved amidoligase_2 domain Amidoligase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 159</td>
<td>17.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 71</td>
<td>7.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 252</td>
<td>28.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Hypothetical protein ORF007** *(Pseudomonas phage PA11)*

**gp1**

<table>
<thead>
<tr>
<th>gp1</th>
<th>+ 73</th>
<th>8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 77</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 60</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 126</td>
<td>14.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conserved ATP-grasp domain**

**ATP-grasp protein**

<table>
<thead>
<tr>
<th>gp1</th>
<th>+ 73</th>
<th>8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 77</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 60</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 126</td>
<td>14.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conserved AIG-2 like domain AIG-2 protein**

<table>
<thead>
<tr>
<th>gp1</th>
<th>+ 73</th>
<th>8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 77</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 60</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 126</td>
<td>14.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conserved YtfP/UPF0131 protein** *(enterobacterial phage φEco32)*

<table>
<thead>
<tr>
<th>gp1</th>
<th>+ 73</th>
<th>8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 77</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 60</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 126</td>
<td>14.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conserved YtfP/UPF0131 protein** *(enterobacterial phage φEco32)*

<table>
<thead>
<tr>
<th>gp1</th>
<th>+ 73</th>
<th>8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 77</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 60</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+ 126</td>
<td>14.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp1</td>
<td>+</td>
<td>97</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>gp1</td>
<td>+</td>
<td>82</td>
</tr>
<tr>
<td>gp1</td>
<td>+ 532</td>
<td>60.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>87</td>
</tr>
<tr>
<td>gp2</td>
<td>+ 161</td>
<td>19.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>67</td>
</tr>
<tr>
<td>gp2</td>
<td>+ 629</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>69</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>270</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>215</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>117</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp2</td>
<td>+</td>
<td>129</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>+</td>
<td>256</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

49
<table>
<thead>
<tr>
<th>gp3</th>
<th>+</th>
<th>115</th>
<th>12.8</th>
<th>Hypothetical protein ViPhICP2p65 (<em>Vibrio</em> phage ICP2)</th>
<th>1e-25</th>
<th>Transcriptional repressor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Transcriptional repressor (<em>Pseudoalteromonas</em> phage H105/1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>+</td>
<td>63</td>
<td>7.2</td>
<td>Conserved poly(A) polymerase head domain</td>
<td></td>
<td>Poly(A) polymerase</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>+</td>
<td>191</td>
<td>21.6</td>
<td>Conserved poly(A) polymerase head domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>+</td>
<td>62</td>
<td>7.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>−</td>
<td>608</td>
<td>64.9</td>
<td>Conserved phage tail repeat-like domain</td>
<td></td>
<td>Possible tail fiber</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collagen triple helix repeat protein (<em>Victivallis vadensis</em> ATCC BAA-548)</td>
<td>8e-06</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Putative tail fiber (<em>Acinetobacter baumannii</em> 1656-2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>−</td>
<td>102</td>
<td>11.5</td>
<td>Hypothetical protein (<em>Vibrio</em> phage ICP2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>-</td>
<td>118</td>
<td>13.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp3</td>
<td>-</td>
<td>144</td>
<td>15.4</td>
<td>ORF-B (predicted transmembrane helix, suspected holin; VpV262)</td>
<td>8e-29</td>
<td>Transmembrane helix (Psipred)</td>
</tr>
<tr>
<td>gp3</td>
<td>-</td>
<td>684</td>
<td>73.7</td>
<td>24/25 41/44</td>
<td></td>
<td>Coiled coils (COILS)</td>
</tr>
<tr>
<td>gp4</td>
<td>-</td>
<td>1413</td>
<td>156.4</td>
<td>38/42 42/47</td>
<td>ORF-U (VpV262)</td>
<td>2e-34</td>
</tr>
<tr>
<td>gp4</td>
<td>-</td>
<td>153</td>
<td>16.0</td>
<td>1/2 9/19</td>
<td>gp26 (Roseobacter phage SIO1)</td>
<td>2e-13</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>-</td>
<td>98</td>
<td>11.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>-</td>
<td>741</td>
<td>82.0</td>
<td>9/11 16/19</td>
<td>Hypothetical protein ORF038 (Pseudomonas phage PA11)</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp27 (Roseobacter phage SIO1)</td>
<td>8e-74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF-P (VpV262)</td>
<td>3e-24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>-</td>
<td>64</td>
<td>6.9</td>
<td></td>
<td>Hypothetical protein ORF038 (Pseudomonas phage PA11)</td>
<td>9e-05</td>
</tr>
<tr>
<td>gp4</td>
<td>251</td>
<td>29.2</td>
<td>3/4</td>
<td>17/27</td>
<td>Hypothetical protein ORF040 (<em>Pseudomonas</em> phage PA11)</td>
<td>1e-109</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>-------</td>
<td>-------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>gp28 (Roseobacter phage SIO1)</td>
<td>7e-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>114</td>
<td>12.0</td>
<td>2</td>
<td>41</td>
<td>Head fiber protein (<em>Bacillus</em> phage B103)</td>
<td>6e-05</td>
</tr>
<tr>
<td>gp4</td>
<td>150</td>
<td>15.4</td>
<td>5/3</td>
<td>36/25</td>
<td>Putative tail fiber, β-structure (Psipred), right-handed β-helix (BetaWrapPro)</td>
<td></td>
</tr>
<tr>
<td>gp4</td>
<td>312</td>
<td>34.7</td>
<td>25/5</td>
<td>92/32</td>
<td>Hypothetical protein ORF044 (<em>Pseudomonas</em> phage PA11)</td>
<td>2e-137</td>
</tr>
<tr>
<td>gp4</td>
<td>264</td>
<td>29.1</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF045 (<em>Pseudomonas</em> phage PA11)</td>
<td>1e-27</td>
</tr>
<tr>
<td>gp30 (Roseobacter phage SIO1)</td>
<td>6e-09</td>
<td>Coiled coil (COILS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>82</td>
<td>9.6</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF046 (<em>Pseudomonas</em> phage PA11)</td>
<td>4e-10</td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>592</td>
<td>67.5</td>
<td>20/19</td>
<td>33/39</td>
<td>Hypothetical protein ORF047 (<em>Pseudomonas</em> phage PA11)</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ORF-H (VpV262)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>gp3, gp5 (<em>Roseobacter</em> phage SIO1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>gp8 (<em>Phormidium</em> phage Pf-WMP3)</td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>535</td>
<td>61.6</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF049 (<em>Pseudomonas</em> phage PA11)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ORF-G (VpV262) 1e-113</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>gp1, gp2, gp4 (<em>Roseobacter</em> phage SIO1) 1e-97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Terminase (Pf-WMP3) 6e-16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Terminase (enterobacterial phage T7) 4e-06</td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>165</td>
<td>19.5</td>
<td></td>
<td></td>
<td>Hypothetical protein ORF050 (<em>Pseudomonas</em> phage PA11)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>636</td>
<td>68.1</td>
<td>21/28</td>
<td>53/66</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>191</td>
<td>21.1</td>
<td>1</td>
<td>11</td>
<td>Hypothetical protein ORF052 (<em>Pseudomonas</em> phage PA11)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>–</td>
<td>68</td>
<td>8.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>-</td>
<td>97</td>
<td>11.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>+</td>
<td>210</td>
<td>23.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp5</td>
<td>+</td>
<td>76</td>
<td>8.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td>+</td>
<td>116</td>
<td>13.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td>+</td>
<td>69</td>
<td>7.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td>+</td>
<td>58</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td>+</td>
<td>158</td>
<td>18.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 3  | Conserved hypothetical protein Acear_1218 3e-21
     | (Acetohalobium arabaticum DSM5501)
<p>| gp6 |  + |  74  |  8.2  |
| 4  |    |    | |
| gp6 |  + |  73  |  8.4  |
| 5  |    |    | |
| gp6 |  + |  125 |  14.2 |
|
|   |   |   |   |</p>
<table>
<thead>
<tr>
<th>gp6</th>
<th></th>
<th></th>
<th>Conserved hydrolase_2 domain</th>
<th>Cell wall hydrolase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>121</td>
<td>14.3</td>
<td></td>
</tr>
<tr>
<td>gp6</td>
<td></td>
<td>144</td>
<td>16.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>152</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>gp7</td>
<td></td>
<td>503</td>
<td>56.2</td>
<td>Hypothetical protein ORF070 (Pseudomonas phage PA11) 3e-03</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.4 Results and discussion

3.4.1 CW02 is a member of the T7-like phage supergroup

The genome of phage CW02, determined by 454 pyrosequencing, had an average read length of 405.47 bp with 44,463 reads and a contig length of 49,391 bp. This corresponded to 366-fold coverage of the genome. The CW02 genome consists of linear dsDNA with a GC content of 47.67%. Analysis of the sequence suggests the genome contains at least 70 ORFs (Table 3.1) and a single tRNA (Arg; anticodon = AGA). Thirty-six of the ORFs appear to be unique to CW02, as PSI-BLAST analysis did not reveal significant matches.

The CW02 genome is similar to the genomes of T7-like phages. PSI-BLAST analysis of putative CW02 ORFs resulted in matches to T7-like phages, including Pseudomonas phage PA11, roseophage SIO1, cyanophage Pf-WMP3, and vibriophage VpV262. The genomes of CW02 and PA11 have at least 24 homologous ORFs in common (Table 3.1), though the functions of the PA11 gene products have not yet been determined. A gene for a T7-like DNA polymerase (gp23) was identified in the CW02 genome by PSI-BLAST analysis,

Figure 3.1: The conserved structure module arrangement of the T7 supergroup. Genomic segments of phages T7, VpV262, Pf-WMP3, and CW02 are shown. The division of the T7 supergroup (dashed line) is characterized by inverted directionality of the genes in T7 compared to the VpV262-like cluster, as discussed. Numbering indicates base pair positions from the 5' to the 3' end of the coding strand. Scale bar, 1 kb.
suggesting an evolutionary relationship between CW02 and T7. However, the genus definition of “T7-like phages” includes the presence of an RNA polymerase encoding gene,\(^{139}\) which was not detected in the CW02 genome. The genomes of SIO1 and VpV262 also lack an RNA polymerase gene.\(^{113}\) Another attribute of the “T7-like phage” genus is the presence of terminal repeats on either end of the viral genome,\(^{139}\) though SIO1 lacks terminal redundancy.\(^{112}\) The CW02 genome also lacks terminal redundancy. The discrepancies of VpV262 and SIO1 with the

![Figure 3.2: Protein composition of mature wild-type CW02 particles as determined by SDS-PAGE and identified by LC-MS. gp39 and gp43 have similar sizes and were not separated in the gel. See Table 3.1 for properties of gene products. Lane MW contains molecular size markers. Peptide fragments from four unique ORFs (gp41, gp46, gp47, and gp55) were detected in the two lowest MW bands (asterisk).](image-url)
definition of the “T7-like phage” genus led to the proposal of a broader, ancestral T7-like supergroup that predates the divergence of phages with and without an RNA polymerase gene, here referred to as “T7-like viruses” and “VpV262-like viruses,” respectively.

The arrangement of structural genes suggests that CW02 belongs to the VpV262-like cluster within the T7 supergroup. A characteristic feature of the T7 supergroup is the presence of a conserved module of late structural genes involved in the assembly of the capsid head. The head structure module of the VpV262-like cluster is arranged such that the genes encoding the terminase, portal, scaffold, and capsid proteins are consecutively ordered and separated by few, if any, gaps (Figure 3.1). The tail structure module is generally encoded leftward of the head module. The structure module of T7-like viruses is related to the VpV262-like viruses by inversion, with the portal, scaffold, and capsid proteins sequentially ordered, though the terminase ORF is rearranged. In general, the genomes of T7-like viruses contain early (e.g., replicative genes) and late genes along the same strand, whereas the replicative and structural genes of VpV262-like viruses are on opposite strands. Bioinformatic analysis of the CW02 genome revealed the same genomic arrangement as other VpV262-like viruses, including VpV262 and Pf-WMP3—a conserved terminase (gp52), portal (gp51), scaffold (gp49), and capsid protein (gp48) closely grouped in sequential order (Figure 3.1) and in the order opposite from that of replicative genes, including the putative DNA polymerase and helicase genes (Table 3.1). Several large ORFs (gp35 to gp47) are found leftward of the head module, and by virtue of the module arrangement and sequence analysis, these genes may be involved in tail assembly (Table 3.1). Based on these observations, CW02 likely belongs to the VpV262-like cluster within the T7 supergroup.
3.4.2 Proteins in the mature CW02 particle

Twelve unique proteins were identified in the mature CW02 virion (Table 3.1). Proteins associated with the virion were analyzed by liquid chromatography (LC)-MS following the excision of detected SDS-PAGE bands. Ten bands were excised and then treated by in-gel trypsin digestion (Figure 3.2). The gel slice corresponding to the lowest-molecular-weight proteins revealed peptide fragments from four unique ORFs (gp41, gp46, gp47, and gp55). Ten of the 12 proteins identified by gel-excision LC-MS were further confirmed by LC-MS of whole, non-SDS-PAGE-treated virions (Table 3.1). Three of the gel bands were identified as the same protein (gp48), and one gel band revealed two proteins of similar sizes (gp39 and gp43). Five proteins present in the virion (gp39, gp41, gp47, gp54, and gp55) did not have detectable homologous counterparts via PSI-BLAST analysis, and their functions remain unknown (Table 3.1).

Proteins in the CW02 head structure module. We identified gp48 as the major capsid protein because gp48 was clearly the most abundant CW02 protein identified by SDS-PAGE (Figure 3.2). PSI-BLAST analysis revealed that gp48 shows some similarities to the putative P22-like coat proteins of Clostridium phages CP26F and 39-O, although the E values were only on the order of $10^{-2}$ (Table 3.1). Pairwise sequence alignment of gp48 and the major capsid protein of T7 (gp10) showed just ~18% identity (55 of 312 residues). Nevertheless, PSI-BLAST analysis revealed that gp48 shows more significant similarities to the putative capsid protein of VpV262 (ORF-K; E value, 3e^{-31}).

Despite poor pairwise sequence alignments, secondary structure prediction methods suggested that gp48 is related to capsid proteins of the T7-like supergroup and other tailed dsDNA phages.
in general. We used the Phyre Protein Fold Recognition Server,\textsuperscript{122} which predicted that gp48 adopts the same fold as the major capsid proteins of HK97 and P22 (P = 0.05). This prediction is significant because pairwise alignment of the HK97 and P22 capsid protein sequences showed just 17% identity. Pairwise alignments of gp48 with the major capsid proteins of HK97 and P22 were similar—just 13% and 20% sequence identities, respectively. This is consistent with observations that the HK97-like fold is more conserved than its sequence.\textsuperscript{105–108} Based on sequence comparisons, we propose that gp49 encodes the CW02 scaffolding protein. gp49 lies in the predicted position of the scaffold protein in the head structure module (Figure 3.1). PSI-BLAST analysis of gp49 identified similarities to other putative scaffold proteins (Table 3.1). Many viruses require scaffolding proteins to direct the proper assembly of virus capsids. The proteins are integral components of phage proheads but are not detected in the mature phage particle.\textsuperscript{141} As expected with phage scaffold proteins, gp49 was not detected by LC-MS of mature virions. Secondary structure prediction algorithms predicted gp49 to consist of an

![Figure 3.3: Coiled-coil prediction profiles of CW02 gp49 and the scaffold protein of phage T7 (gp9). The probabilities of a coiled coil in either protein were determined using the program COILS.](image)
entirely helical structure (data not shown), which is consistent with the crystal structure of phage 29 scaffolding protein. Although the sequences of gp49 and the scaffolding protein of bacteriophage T7 share only ~15% sequence identity (41 of 264 residues), both proteins are predicted to have two or more regions that can form coiled-coil arrangements (Figure 3.3). The predictions differ in length and probability, but they suggest that the two proteins have structural similarities.

Based on sequence comparisons, CW02 gp51 appears to be a portal protein. Tailed dsDNA phages encode portal proteins that enable DNA passage during packaging and ejection. The portal also functions as a connector between the capsid and tail proteins. PSI-BLAST analysis of gp51 showed a conserved domain belonging to the head-tail connector superfamily. The sequence of gp51 is strongly conserved with the putative portal proteins of VpV262, SIO1, and Phormidium phage Pf-WMP3/4 (Table 3.1). We used secondary structure prediction methods to investigate possible similarities between gp51 and the portal protein of phage T7 on the structural level and found the profiles to show some resemblance (data not shown).

Sequence comparisons suggest that CW02 gp52 is the terminase protein. Phage terminases are enzymes that facilitate packing of the viral genome into the phage proheads and, like scaffold proteins, are not found in the mature virion. As expected, gp52, was not detected by LC-MS of mature CW02 particles. CW02 gp52 had weak sequence similarity to the terminase protein of phage T7 but stronger sequence similarity to the putative portal proteins of VpV262 and SIO1. Secondary structure prediction of gp52 showed an arrangement similar to that of the terminase of phage T7 (data not shown), suggesting that their functions are conserved.
3.4.3 Proteins in the CW02 tail structure module

Putative tail proteins were observed leftward of the head structure module. This placement is consistent with tail genes in the known T7-like supergroup. A number of large ORFs are found in an appropriate location of the CW02 genome that could reasonably be thought to encode tail-associated proteins. For instance, PSIBLAST analysis of gp35 revealed a conserved

---

**Figure 3.4:** Electron microscopy of CW02 by negative stain (A) and cryogenic (B) methods. Stubby or thin tail-like features are marked by white arrowheads. Collapsed, empty particles are shown by black arrowheads. The inset in panel B shows examples of particle images used in computing the 3D reconstruction (black scale bar, 50 nm).
“phage tail-like repeat” domain. Also, gp40, encoded by the largest putative ORF in the CW02 genome, is a 1,413-amino-acid protein—consistent with the large sizes of phage tail sheath proteins. Trimeric structures of coiled coils and β-helices are a common motif in tail-associated proteins, such as the tail spike protein of P22 (homotrimer of β-helices)\(^{143}\) and tail fibers of bacteriophage T7 (triple-stranded coiled coils).\(^{144}\) The presence of coiled coils and β-helices in candidate ORFs within the putative tail structure module was examined. The BetaWrapPro program revealed right-handed-helical motifs in gp43 and gp47, whereas the COILS program detected coiled-coil motifs in gp39 and gp41.

![Figure 3.5: Cryo-EM-based reconstruction of CW02. (A) Stereo image of the CW02 capsid surface rendered at a 1-σ contour level. Structures are shaded from light to dark, corresponding to the radius. (B) The central slice of the reconstruction (perpendicular to a 2-fold symmetry axis). Black represents protein or DNA density. Bar, 25 nm. Black arrowheads indicate turrets. The arrow indicates potential capsid-dsDNA contacts. (C) Plot of average density with respect to the radius of the reconstruction. DNA and capsid density peaks are labeled.](image-url)
3.4.4 Electron microscopy of CW02 particles

CW02 virions are morphologically similar to members of the family *Podoviridae*. Negatively stained CW02 particles were visualized by transmission electron microscopy and revealed particles with icosahedral heads and a short, tail-like feature (Figure 3.4A). However, the tail-like structures showed inconsistencies among individual particles. For example, a short, stubby tail was observed in some of the particles, whereas in other particles, a short, thin, fiber-like feature was more apparent. Cryo-EM was used to visualize CW02 particles in the solution state (Figure 3.4B). Although similar stubby and thin-fiber tails were observed, tail-like structures were more difficult to discern in cryo-EM micrographs, which may be a consequence of decreased contrast in samples with high salt concentrations.\(^{145}\)

3.4.5 T = 7 levo icosahedral capsid lattice

Cryo-EM images of CW02 were used to compute a 3D reconstruction. The 3D reconstruction of the CW02 head was determined to a 16-Å resolution using 8,695 particle images (out of a total of 10,677 particle images) and applying icosahedral averaging to maximize the signal-to-noise ratio (Figure 3.5A).

The head of CW02 is arranged in a T = 7 icosahedral lattice (Figure 3.5A). T = 7 lattices have two possible arrangements that are mirror images of each other. Because transmission electron microscopy images are 2D projected views, information about the object handedness is lost in its 3D reconstruction. The structures of all currently known T = 7 bacteriophages, including lambda\(^{146}\), HK97\(^{104}\), and T7\(^{105,147,148}\), are of the levo orientation. We performed tilt experiments to determine the handedness of the CW02 capsid. This method relies on the cross-correlation of images of tilted specimens with projections of the tilted model of either handedness.\(^{132}\) Like that of the other T = 7 phages, the handedness of CW02 was determined to be of the levo orientation.
For particles showing a clear difference between left- and right-handed T = 7 lattices (i.e., correlation coefficient difference of ≥ 0.10), the average correlation coefficient was 0.41 ± 0.06 for T = 7 levo handedness and 0.26 ± 0.06 for T = 7 dextro handedness (n = 40).

Figure 3.6: Rigid-body fit of the HK97-like fold into the CW02 capsid. (A) Ribbon diagram of the HK97 protein with pertinent helices and domains labeled. Yellow, N terminus; green, cross-linking arm; blue, A domain; red, P domain. Yellow, green, light red, and light blue indicate coordinates omitted from the CW02 fitting. (B and C) Frontal (B) and side (C) views of the HK97 A and P domains (chains A to G) fitted into CW02 hexamers (purple, red, yellow, green, teal, and blue, chains A to F, respectively) and pentamers (orange, chain G). Each subunit is colored according to conventions used previously. Solid lines in panel B outline the wedge-like conformation of the A domain. (D) Central section of the CW02 reconstruction fitted by multiple copies of the HK97-like fold. The white arrowhead points to concentric layers of the DNA genome. In all of the panels; representative gpD-like densities are indicated by stars; the A and P domains are labeled in italics; symmetry and hexamer axes (dashed lines) are labeled with pentagons (5-fold), triangles (3-fold), ovals (2-fold), and a hexagon (local 6-fold, in panel C only); and turret densities were omitted.
3.4.6 Architecture of the CW02 head

The diameters of the CW02 head measured 691 Å, 627 Å, and 585 Å along the icosahedral 5-fold, 3-fold, and 2-fold axes, respectively. The average thickness of the capsid shell is approximately 37 Å. Sixty hexamers and 12 pentamers comprise the icosahedral head.

We postulate that the CW02 capsid has a conserved structure with that of the T7 capsid and other phages of the Caudovirales order. To date, the capsids of all tailed dsDNA bacteriophages appear to adopt the HK97-like fold. For instance, the pseudoatomic capsid structure of phage T7, solved by cryo-EM and modeling methods, showed the conserved fold.\textsuperscript{105,147} The centers between two hexameric capsomers of the CW02 reconstruction are separated by approximately 140 Å, which is a characteristic feature of the HK97-like fold.\textsuperscript{106,149,150} The fold consists of two domains: the axial (A) domain and the peripheral (P) domain (Figure 3.6A).\textsuperscript{104,134} The A domain is a compact, wedge-like structure with two α-helices (α5 and α6) and a four-stranded β-sheet. The P domain is elongated, consisting of a long helix (α3) accompanied by a long β-sheet. Together, the domains form an L-shaped protein with a continuous hydrophobic core.\textsuperscript{134} The HK97 capsid protein also bears an extended amino-terminal arm and a cross-linking loop (Figure 3.6A), but these segments are not as conserved in other tailed dsDNA phages.

To further investigate the similarities between CW02 and HK97-like capsids, we performed rigid-body fitting of the HK97A and P domains into the CW02 reconstruction (Figure 3.6B to D). The Ca backbones of the A and P domains were fitted into the density map as either hexamers or pentamers. Flexible loops were omitted from the fitting, leaving 128 to 130 residues in the fitted core structure. Similar results were obtained by fitting hexameric and pentameric capsomers, e.g., the agreement between the characteristic slants of the α5 and α6
helices into the cryo-EM densities (Figure 3.6B, solid lines). The elongated α3 helix and accompanying P domain β-sheet also fit snugly into the CW02 densities (Figure 3.6C). The P domains interact in a 3-fold fashion at the 3-fold and quasi-3-fold regions. Our fit showed agreement between the P domain coordinates and the CW02 map, though voluminous triangular densities were found above each of the P domain fits (Figure 3.6B and C, star). The overall fit is consistent with previous studies that showed the intracapsomere stabilization role of the A domain and the intercapsomere contacts made by the P domain.\textsuperscript{134}

Other than the main fold, other elements of the HK97 capsid protein appear to be missing from CW02. The HK97 capsid protein structure consists of a cross-linking loop and extended N-terminal arms that stabilize interactions within each capsomere.\textsuperscript{104,151} No evidence of cross-linking is seen in CW02 due to the absence of gp48 oligomers in SDS-PAGE analysis (Figure 3.2). The HK97 cross-linking loop and N-terminal arms fit poorly into the intercapsomere space, suggesting that CW02 has different means of intercapsomere contacts (data not shown).

3.4.7 gpD-like densities at the 3-fold and quasi-3-fold axes
Other densities in the CW02 reconstruction were not accountable by the HK97-like fold. Most prominently, the intersection of every three capsomeres is bridged by a triangular body found on the 3-fold and quasi-3-fold axes (Figure 3.5 and stars in Figure 3.6B and C). Both types of triangular densities showed consistent structures, despite icosahedral averaging, suggesting that the proteins on the 3-fold and quasi-3-fold axes are identical.

The functions of the triangular densities are unknown, but their structures are reminiscent of the auxiliary protein gpD trimer of bacteriophage lambda (Figure 3.7).\textsuperscript{107,146} Similar to the gpD trimer, the positions of the triangular densities in CW02 on the 3-fold and quasi-3-fold axes
suggest an intercapsomere stabilization role\textsuperscript{107} and its positions coincide with the covalent cross-links formed by the HK97 capsid protein.\textsuperscript{104} The P domains of the HK97-like fold form intercapsomere contacts directly beneath the triangular densities (Figure 3.6). In other HK97-like phages, intercapsomere interactions by the P domain are accompanied by more extensive intercapsomere interactions from additional accessory proteins (including gpD of lambda\textsuperscript{146} and hoc of T4\textsuperscript{152}), covalent cross-links (as in HK97\textsuperscript{104}), or additional domains as part of the capsid protein. For instance, the pseudoatomic structure of bacteriophage Φ29 showed that the major capsid protein consists of the HK97-like domain and an additional immunoglobulin-like domain.\textsuperscript{108} Bacteriophage P22 also consists of extra globular densities on the outermost surface of the capsid.\textsuperscript{106} However, unlike the densities seen in CW02, the additional domains in Φ29 and P22 do not form trimeric structures. The triangular densities likely belong to a separate domain of CW02 gp48 and are not a separate, minor capsid protein. If the triangular densities of CW02 represent separate protein subunits, then 405 to 420 copies of the protein would be required per capsid (i.e., the same number of major capsid protein subunits), as is the case with gpD of lambda.\textsuperscript{107} Yet, SDS-PAGE analysis of mature CW02 particles did not reveal other proteins as abundant as gp48, arguing against the presence of a separate gpD-like protein (Figure 3.2). BLAST analysis of the CW02 genome did not reveal the presence of a gene like that for gpD. Thus, the triangular densities of CW02 most likely represent an additional, non-core domain of gp48, similar to the amino-terminal arm and cross-linking domain in HK97 (Figure 3.6A),\textsuperscript{104,134} the Ig-like domain of Φ29,\textsuperscript{108} and the additional globular domains of P22.\textsuperscript{106} The additional ~180 CW02 gp48 residues unmodeled in our fitting not only connect pieces of the core domain but also likely form the additional gpD-like domain observed above the fitted core HK97 structure (Figure 3.6).
3.4.8 Lack of tail density in 3D reconstructions

The presence of a tail structure module in the CW02 genome (Table 3.1) and apparent tail-like features in the EM images (Figure 3.4) suggests that each capsid is associated with a short tail. From these images, the tail appears to be a stubby cylindrical component centered on a thin, short rod.

![Segmented views of the subunits that make up the CW02 (A) and lambda\textsuperscript{107} (B) capsids. Representative hexameric and pentameric capsomer subunits of the asymmetric unit are colored as in Figure 3.6B. Representative trimeric, gpD-like densities at the intersection of every three capsomers are pink. The turret density of CW02, located along the 5-fold axis, is brown.](image)

Yet, despite the use of asymmetry or cyclic 5-fold (C5) symmetry during the reconstruction process, none of our reconstructions produced a tail-like structure. The addition of artificial, cylindrical, tail-like (short and stubby or short and thin) or portal-like densities to a single vertex of the CW02 capsid as an orientation-refining fiducial marker did not produce different structures (data not shown). One noteworthy difference between the pentameric and hexameric capsomeres is the presence of an exterior bulge stemming from the center of the pentamers (Figure 3.5). The bulge may correlate with an additional protein complex and may be part of the tail apparatus. Alternatively, the tail-like material observed in negative stain and cryo-EM images (Figure 3.4) may be partial discharge of DNA or discharge of protein from the capsid. In
many of our preparations of negatively stained and frozen-hydrated particles, tail-like densities were not observed in the electron micrographs (data not shown). Nevertheless, considering all of the evidence, we think that CW02 most likely contains a tail that was “hidden” in our cryo-EM 3D analysis by the decreased contrast of the high-salt buffer.145

3.4.9 Structure of the encapsidated dsDNA and lack of internal membrane
HK97-like phages appear to have conserved mechanisms of dsDNA packaging, such that the viral genome is cylindrically spooled within the capsid interior.153,154 We observed at least eight concentric layers of densities within the icosahedrally averaged head of CW02 (Figure 3.5C and D). Each layer is spaced approximately 24 Å apart, similar to the dsDNA densities observed in other bacteriophage structures, including T7148,153 and lambda.154 The layering pattern is consistent with the DNA spooling model established for T7.153 Weak densities are also observed at the interface between the outermost DNA layer and the capsid (Figure 3.5C), indicating possible interactions between the capsid protein and dsDNA core.

Some turreted, extremophilic viruses of the PRD1-like lineage, including Sulfolobus turreted icosahedral virus and SH1, encase an internal lipid membrane that is thought to aid in the injection of the genome into the host cell.155 Reconstructions of CW02 did not show any evidence of material other than dsDNA within the capsid.

3.4.10 Presence of turrets on 5-fold vertices
On each icosahedral vertex, the CW02 reconstruction showed the presence of turret-like structures that extend ~40 Å beyond the capsid head. The turret densities appear disconnected from the capsid, suggesting that the linker between the turret and capsid is disordered or incompatible with 5-fold symmetry. To investigate whether the turrets are unique to a single
vertex or if they are positioned at each of the 12 5-fold vertices of the capsid, we computed reconstructions of CW02 without any imposed symmetry (in both the orientation finding and 3D image reconstruction processes). These reconstructions also showed the presence of disconnected turrets on all vertices (data not shown). The 3D reconstructions and class averages from the turrets (see Materials and Methods) showed inconclusive results, but control experiments with images of poliovirus complexed with antibodies to the 5-fold region (J. Lin et al., unpublished data) produced correct results. Therefore, CW02 turrets appear to lack a consistent order. The function of the CW02 turrets and the gene encoding the turret protein remain unknown, but the turrets may be unique appendages similar to those seen in other extremophilic viruses, including the archaeal viruses Sulfolobus turreted icosahedral virus\textsuperscript{155,156} and halovirus SH1.\textsuperscript{145} Other icosahedral viruses with protein turrets or spikes on vertices include adenovirus\textsuperscript{157} and bacteriophage PRD1.\textsuperscript{158} The function of turrets is thought to be to assist in host adsorption.\textsuperscript{145} Collectively, turreted viruses belong to the PRD1-like lineage and consist of capsids with a single or double β-jellyroll fold, unlike CW02. Thus, the presence of turrets on CW02 could represent structural features conserved between viruses of the PRD1-like and HK97-like lineages.
Chapter 4  Resolving double disulfide bond patterns in SNAP25B using liquid chromatography–ion trap mass spectrometry

Author’s Note: This chapter introduces a novel structural analysis of a complex disulfide bridged peptide sequence, capable of elucidating disulfide bonding patterns. I assisted in algorithm development, mass spectrometry, and document preparation. These results were published in the Journal of Mass Spectrometry.

4.1 Abstract

Complex disulfide bond patterns in synaptosomal-associated protein of 25 kD B (SNAP25B) are thought to regulate neurotransmitter release in response to oxidative stress. However, the steric feasibility of each possible disulfide pattern in SNAP25B has not been assessed. To assess the steric feasibility of hypothesized closely spaced complex disulfide patterning in SNAP25B and also the feasibility of identifying complex disulfide bond patterns with MS, we have developed a novel probabilistic analysis to unambiguously resolve complex double disulfide bond patterns by using an ion trap mass spectrometer. We analyzed fragmentation patterns of singly linked peptides to determine likely fragmentation events in an ion trap mass spectrometer and observed double and single backbone cleavage along with heterolytic cleavage of the disulfide bond. We modeled these same events in the doubly disulfide linked SNAP25B peptide and used a cumulative hypergeometric distribution with top–down scoring to both identify and differentiate these bonding patterns. Because of the presence of unique MS/MS peaks, two of the bonding patterns were directly identified. The third was assigned on the basis of full chromatographic separation and confirmed by modeling triple breakage fragments. In total, this work

Publication authors are Ogawa N, Taylor RM, Woodbury DJ, Prince JT.
demonstrates the feasibility – and also limitations – of identification of complex intradisulfide patterns by using ion trap-based collision-induced dissociation-based fragmentation methods.

4.2 Introduction

Disulfide bonds stabilize native protein conformation but are hypothesized to function as a redox buffer and sensor of oxidative stress conditions. At the synapse, neurotransmitter release is mediated by soluble N-ethylmaleimide (NEM)-sensitive fusion factor attachment protein (SNARE) proteins. In the presence of calcium ions, SNARE proteins assemble into a catalytic SNARE complex and stimulate the release of neurotransmitters. Under oxidative conditions, the SNARE complex is thought to become dysfunctional, contributing to symptoms of neurodegenerative disease such as Parkinson’s and Alzheimer’s. One of the SNARE proteins, synaptosomal-associated protein of 25 kD B (SNAP25B), is a known target of oxidation, and oxidation of SNAP25 inhibits the formation of the SNARE complex.

The regulation of SNARE complex formation by SNAP25 is likely controlled by the oxidation of a regulatory sequence composed of four closely spaced cysteine residues, by some combination of differential palmitoylation and oxidation. Complete oxidation results in one of the three possible configurations (Figure 4.1) for any peptide containing four cysteines. Molecular dynamic simulation has shown that the inhibition of the SNARE complex could be caused by a
change in structural conformation of SNAP25 via formation of disulfide bonds among the four cysteine residues in SNAP25. These cysteines of SNAP25 are established to be a site of differential palmitoylation and known to play a role in subcellular localization and membrane anchoring of SNAP25. These competing modifications suggest that identification and characterization of disulfide bond patterns can be crucial to understanding the regulation of SNARE complex formation.

Current methods for disulfide bond order determination are fairly limited. Protein structure determination using NMR can be used to identify the location of specific disulfide bonds but requires a large amount of pure sample and is therefore prohibitive for complex mixtures. MS-based methods are ideal for isolating and characterizing samples in a complex mixture. The typical protocol involves a two-step tagging process, separated by a reduction step. With this method, unbound cysteines are labeled with one irreversible alkylating reagent (typically N-ethylmalamide), and the other pair of bonded cysteines is reduced and then labeled by another irreversible alkylating reagent (typically iodoacetamide). Cysteines modified with the same compound are deductively found to be paired. While effective for single disulfide bonds, this method fails to elucidate the bonding pattern if multiple disulfide bonds are present. The identification of the pairing pattern has been accomplished, however, through clever application of the reduction/alkylation methodology by stepwise methods where cysteines are partially reduced and then differentially alkylated after each reduction step. Although effective for identifying complex disulfide bond pairing patterns, the method has drawbacks: it is not a direct detection of disulfides and thus cannot differentiate disulfide bonds from other oxidative
modifications; it is not effective when the disulfide bonds have the same reduction potential; and it still requires a large amount of protein.\textsuperscript{167,168}

An alternative approach to the partial reduction protocol is nonreduced/nonalkylated mass spectrometric analysis of disulfide bond patterns. This method relies on cleavage of both the disulfide bonds and opening of the disulfide loop by breaking internal peptide bonds through collision-induced dissociation (CID).\textsuperscript{169–171} This technique is particularly effective when proline is present within the disulfide loop because of increased fragmentation efficiency at X-Pro linkages.\textsuperscript{160} Excellent tools to facilitate identification of simple disulfide-bonded peptides have been introduced, such as MassMatrix, but are unable to search more complex disulfide patterns that require opening of the disulfide loop for identification.\textsuperscript{172}

Most previous disulfide bond linkage identification efforts were performed in collision cells at higher energies\textsuperscript{171,173} to enable amino acid sequencing within the disulfide loop region. Recently, a group used ion trap-induced CID to successfully generate internal fragments in a toxin that contains disulfide bonds\textsuperscript{62} and characterize a triple disulfide bond-containing peptide toxin. Unlike collision cells, fragmentation within an ion trap is expected to result in double or even triple cleavage products at normal energies because cyclic cleavage products experience sustained excitation until fragments dissociate.

Using conotoxins as an example, one group has demonstrated that an identification of complex intradisulfide bonds may be feasible by modeling double backbone cleavage and disulfide bond breakage.\textsuperscript{174} Although this is a significant breakthrough, a method of resolving complex intramolecular disulfide bonds has not yet been achieved – there is no systematic or statistical scheme to unambiguously resolve disulfide bond patterns. The development of a scheme that
convincingly differentiates among possible disulfide patterns is crucial because the intensity of MS/MS spectra from the double backbone cleavage or disulfide bond cleavage is lower than conventional fragmentation ions and therefore difficult to distinguish from the noise. Here, we report CID-based mass spectrometric determination of all three distinct double disulfide bond patterns possible in the cysteine-rich region of SNAP25B by using chromatographic separation and statistical analysis of shared and unique MS/MS peptide fragments, including double fragmentation events. This work suggests the feasibility and also limitations involved in identification of complex intradisulfide patterns by using liquid chromatography (LC) and ion trap mass spectrometers.

4.3 Materials and method

4.3.1 Protein and synthetic peptide preparation

The recombinant SNAP25B clone was expressed as a glutathione S-transferase (GST) fusion protein in Escherichia coli and purified following the published protocol with minor modification. During the induction step, optical density of the media was kept at 0.8~1 A.U. to optimize the protein synthesis for SNAP25B. The final concentration of the purified protein was ~0.1 mg/ml. Identity of SNAP25B was confirmed by Western blot analysis and MS (Polyclonal SNAP25B, Abcam, Cambridge, MA). The SNAP25B plasmid was a generous gift from J. E. Rothman (Memorial Sloan-Kettering Cancer Center, NY). Synthetic peptides of sequence FCGLLVLPCNK and FLGLCVCPLNK were purchased from Genescript Inc. (Piscataway, NJ). These peptides are identical to the tryptic fragment of the cysteine-rich region from SNAP25B except for the replacement of two cysteines with leucines.
4.3.2 Sample preparation

Synthetic SNAP25B peptides were suspended in water and incubated ~1 h in air to oxidize. Peptides were subsequently acidified in 1% formic acid. One picomole of peptide was injected in LC–MS/MS. Purified SNAP25B was run on standard sodium dodecyl sulfate polyacrylamide gel electrophoresis (4–20% gel with antioxidant). The excised SNAP25B band was subjected to a published in-gel tryptic digestion protocol,176 altered by skipping reduction by dithiothreitol and replacing the iodoacetamide solution with 80mM solution of NEM instead. This alkylation step, which does not affect peptides with two disulfide bonds, increased the mass of singly disulfide-bonded peptides to prevent spectral overlap between the single and double disulfide species. Sodium dodecyl sulfate polyacrylamide gel electrophoresis was performed using commercially available equipment and protocols (Pierce, Rockford, IL). Trypsin was purchased from Promega, Madison, WI.

4.3.3 Liquid chromatography and MS

A sample (27 picomoles) was loaded via an Ultra2D UPLC/autosampler system (Eksigent, Dublin, CA) to a 0.17 ml bed volume C-18 stem trap (Optimize Technologies Inc., Oregon City, OR) at 3 ml/min in 95% Buffer A (95% water, 5% acetonitrile, and 0.1% formic acid), 5% Buffer B (0.1% formic acid in acetonitrile) for 10 min. Elution was carried out at 325 nl/min through a 20-cm, 75-mm ID NanoAcquity C18 column (Waters Corporation, Milford, MA) by running from 98 to 88% Buffer A over 5 min, 88 to 65% Buffer A over 120 min, a 5-min ramp to 5% Buffer A for 12 min, and a 4-min ramp to 95% Buffer A for 6 min to equilibrate the column for subsequent runs. Column eluent was ionized via nanospray (2.0 kV, capillary) and analyzed by the LTQ Orbitrap XL (Thermo Fisher Scientific Inc., Waltham, MA). Survey scans were conducted at 60,000 resolution in the Orbitrap mass detector. Peaks were then selected for
further ion trap CID fragmentation if they were on an inclusion list or were of the highest intensity from the preview survey scan. The inclusion list consisted of all possible parent masses for the charge state +2 tryptic fragment representing FCGLCVCPCNK. CID was carried out at 35% normalized collision energy.

4.3.4 Spectra analysis
RAW data files were converted to mzML files via ProteoWizard.26 MS/MS spectra were extracted via mspire.177 We wrote a software package for predicting cross-link cleavage events and analyzing extracted MS/MS spectra (freely available at https://github.com/princelab/double_disulfide_searcher) in the Ruby programming language. An overview of the program is provided in supplemental material (S1). All MS/MS fragmentations were carried out on z = +2 precursor mass. Because of the low intensity of the multiply cleaved peptide peaks, we analyzed only z = +1 peaks for both the MS/MS spectra of trypsin digested SNAP25B and synthetic peptide mutants. The mass tolerance parameter for all the MS/MS experiments was set at 0.3, yielding a bin size of 0.6 Thomson (Th) for both experimental and predicted spectra.

For the analysis of the synthetic peptides FCGLLVLPCNK and FLGLCVCPLNK, experimental MS/MS spectra were averaged for each using Xcalibur (Thermo Fisher Scientific Inc., Waltham, MA). We confirmed that there is no intermolecular disulfide bridge (dimer) between the peptides in our preparation by ensuring that no higher charge states existed at the given Th value. For the peptide FCGLLVLPCNK in the +2 state, there was a complete chromatographic separation between the reduced ((MW + 2)/2) and oxidized (MW/2) forms. For the peptide FLGLCVCPLNK, the chromatogram showed overlap between the reduced and oxidized form.
However, we confirmed the absence of cross contamination between the oxidized and reduced MS/MS through manual examination of the spectra. The reduced species b and y series ions were not observed in the MS/MS spectra of the oxidized species indicating that the 1 Th difference between the reduced and oxidized species was sufficient to resolve the two completely.

We eliminated the possibility of spectral contamination for the tryptic-digested samples by treating them with NEM, which causes a 250-Da shift to the single disulfide species and a 500-

![Figure 4.2](image)

**Figure 4.2:** CID cleavages observed for test peptides. (a) Diagram of possible cleavages. (b) CID cleavage for each possible cleavage type is compared as percent of the possible cleavages. In the longer loop case, C88L C90L SNAP25 peptide, a significant number of double peptide backbone cleavages and heterolytic disulfide cleavages are generated. Random is a pool of 1000 randomly selected masses generated at each Q depth. Q depth between 3 and 10 was typical for later searches, as these optimized the signal to noise ratio. (c) C85L C92L SNAP25B peptide generates far less double and heterolytic disulfide cleavages.
Da shift to the fully reduced species. The double backbone cleavages use b-fragmentation and y-fragmentation schemes described by Hunt et al.\textsuperscript{160} and Clark et al.\textsuperscript{173} Heterolytic and homolytic disulfide cleavages, which generate persulfide/dehydroalanine and cysteine/thioaldehyde, respectively, were predicted on the basis of the mechanism described by Choi et al.\textsuperscript{171}

4.4 Results

![MS/MS spectra of the oxidized peptide FCGLLVLPCNK (C88L C90L mutant). The fragmentation of the peptide with CID shows robust double backbone and heterolytic cleavages. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: y10, z11-CO, y11-CO, and z11. The Q depth of this analysis was set to Q= 10, and details of the peaks are presented in Supplemental Data Table. The +SSH and H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision-induced dissociation (persulfide and dehydroalanine, respectively). Peaks under the bold black lines (×5 and ×2) are magnified as designated.](image)

Figure 4.3: MS/MS spectra of the oxidized peptide FCGLLVLPCNK (C88L C90L mutant). The fragmentation of the peptide with CID shows robust double backbone and heterolytic cleavages. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: y10, z11-CO, y11-CO, and z11. The Q depth of this analysis was set to Q= 10, and details of the peaks are presented in Supplemental Data Table. The +SSH and H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision-induced dissociation (persulfide and dehydroalanine, respectively). Peaks under the bold black lines (×5 and ×2) are magnified as designated.
First, we demonstrate that double backbone and heterolytic disulfide cleavages occur in the ion trap and present a method for deductively resolving the three patterns of double disulfide bonds that are possible in a peptide with four cysteines. We then use this method to confidently identify all three complex disulfide bond patterns of the four cysteine peptides of SNAP25B. Finally, we confirm the identifications with an analysis of triple backbone cleavage fragments.

4.4.1 Double peptide backbone and disulfide cleavages in an ion trap
As a test case to characterize the propensity for double backbone cleavage and disulfide (heterolytic/homolytic) cleavage events in an ion trap (Figure 4.2a), we fragmented two single disulfide peptides, where two cysteines in SNAP25B are substituted with leucines (peptides shown in Figure 4.2b–c). To characterize the mode of fragmentations at different levels of intensity, MS/MS spectra were divided into 100 Th bin, and top N number of highest intensity peaks were extracted from each bin and analyzed (the value of N is referred as Q(queue) depth in this article). This method of MS/MS spectra optimization is used in the proteomic search engine Andromeda.23 The propensities for different cleavage types were monitored by calculating %observed/predicted fragments for each Q depth. The first SNAP25B C88L C90L mutant peptide (with a longer disulfide loop) showed significant double backbone cleavage fragments, with comparable intensity to single cleavage fragment ions (Figs 2(b) and 3 and Supplemental Table). For this peptide, heterolytic disulfide cleavage occurs well above background levels but less commonly than the single or double backbone cleavages. The second SNAP25B C85L C92L mutant peptide (with a shorter disulfide loop) showed less double backbone cleavages and heterolytic disulfide cleavage, with single cleavage being the dominant type (Figs 2(c) and 4 and Supplemental Table). In both cases, homolytic disulfide cleavages followed the same trend as random fragments, suggesting that homolytic disulfide cleavage is not occurring under these conditions.
conditions. The fragmentation pattern observed here is consistent with the study by Choi et al. that reported frequent observation of C–S bond cleavages and rare occurrence of S–S bond cleavage under CID.171

4.4.2 In silico MS/MS fragmentation

![MS/MS spectra of the oxidized peptide FLGLCVCLPPLNK (C85L C92L mutant). The fragmentation of the peptide with CID shows few double backbone and heterolytic cleavages at low intensity. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: a2, a4, y10-CO, z11-CO, y11-CO, b11, and z11. The Q depth of this analysis was set to Q= 10, and details of the peaks are presented in supplemental data table. The +SSH and −H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision-induced dissociation (persulfide and dehydroalanine, respectively).]
Theoretically, the SNAP25B cysteine-rich region may form three different bonding patterns upon double disulfide bond formation. We postulated that double backbone cleavage fragmentations would generate diagnostic fragments unique to a specific disulfide patterns, whereas heterolytic disulfide cleavages would generate fragments that could be employed to give the primary amino acid sequence.

To confirm our hypothesis, in silico MS/MS fragmentation simulations were performed on the SNAP25B cysteine-rich region of the three disulfide bond variants, considering single, double backbone, and double backbone/heterolytic cleavages. The different disulfide patterns of SNAP25B resulted in different and unique MS/MS fragmentation upon single, double backbone, and double backbone/heterolytic cleavages (Table 4.1). However, the cross-over pattern with the double disulfide did not produce unique diagnostic fragments upon in silico MS/MS fragmentation.

Table 4.1: The number of m/z fragments predicted for each fragmentation scenario in the SNAP25B cysteine-rich region. The numbers inside the parentheses are fragments that uniquely occur in each specific disulfide bond pattern. These fragments can be used to detect each disulfide pattern.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Single</th>
<th>Double backbone</th>
<th>Backbone/heterolytic disulfide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>10 (4)</td>
<td>16 (6)</td>
<td>20</td>
</tr>
<tr>
<td>Parallel</td>
<td>6</td>
<td>21 (10)</td>
<td>20</td>
</tr>
<tr>
<td>Cross-over</td>
<td>6</td>
<td>9</td>
<td>20</td>
</tr>
</tbody>
</table>
These patterns are generalizable to all four cysteine-containing peptides: the cross-over pattern will not generate unique fragments when considering single or double fragmentation events and therefore is not uniquely identifiable, unlike the open and parallel patterns. Regardless of the peptide sequence, a crossover pattern will always produce double cleavage fragments that are subsets of the open and parallel patterns. Nevertheless, two approaches, both employed for this project, can be used to ascertain the identity of the cross-over pattern: (1) deductive identification conditioned on chromatographic separation and confident identification of the other patterns and (2) expansion of the fragmentation scheme to include triple backbone cleavages that do produce unique fragments for the cross-over pattern.

4.4.3 Complete chromatographic separation
Triple backbone cleavage had never been previously demonstrated in the literature, and therefore, we developed a scheme to confidently identify all three patterns on the basis of complete chromatographic separation and deductive logic (Figure 4.5). The following conditions must be met to deduce the identity of bonding patterns that yield no unique fragments: (1) the MS/MS fragments must confirm that the peptide is in fact the linked peptide in question

Figure 4.5: Decision tree logic supporting the correct identification of each disulfide pattern.
(i.e., a member of the set of possible bond arrangements and not some other peptide), (2) the various linked peptides must be separated chromatographically, and (3) the other LC peaks and their corresponding patterns must be confidently identified.

4.4.4 Scoring method
Following this logic tree, we developed a software that utilized an identification algorithm based upon the proteomic search engine, Andromeda,\textsuperscript{23} with some modifications to account for the double backbone and heterolytic disulfide cleavages. The observed MS/MS spectrum is compared with the predicted spectrum and assigned a goodness of fit score on the basis of the set of shared fragmentation ions from all possible disulfide bond patterns. The maximal score is calculated from the set of scores obtained from matching the shared predicted fragments with experimental spectra at Q depths of 1–10, where the Q depth that maximizes the score is accepted. The score is calculated using a cumulative hypergeometric distribution statistical model in lieu of the binomial distribution.\textsuperscript{178,179} We generate two scores, (1) a primary sequence score that confirms the primary amino acid sequence of the peptide and (2) a pattern diagnostic score that is used to identify specific disulfide patterns, where \( \text{score} = -10 \times \log_{10}(p\_value) \). The primary sequence score is analogous to a conventional Andromeda score, representing the match quality of the fragment ions to the predicted fragment ions (Table 4.1). The pattern diagnostic score represents the significance of the match to the unique fragment ions of each predicted disulfide pattern (Table 4.1). These two values are used to evaluate the presence of a pattern (Figure 4.5). An overview of the program is provided in Supplemental Material (S1).

To be stringent in identifying the primary amino acid sequence of the peptide, 99,999 decoy fragmentation spectra, composed of 33,333 unique, scrambled FCGLCVCPCNK sequences each
with all three possible disulfide bond patterns represented, were matched against the experimental spectrum. The primary amino acid sequence was validated only when the primary sequence score of the target peptide was statistically significant and the highest among the primary sequence scores of the 99,999 decoy spectra.

4.4.5 Identification of the three different disulfide patterns

The three double disulfide bond patterns were separated and identified from an in-gel tryptic digest of oxidized and affinity purified SNAP25B protein. To prevent spectral overlap between the single and the double disulfide species of SNAP25B, we treated the sample with NEM to tag open sulfhydryls on any single disulfide species. This induces a 250-Da mass shift to the single disulfide species, to prevent spectral overlap between the different oxidation states. This is crucial because the parent mass of the single and the double disulfide species differ by 1 Th at \( z = +2 \), which is sufficient to cause spectral cross-contamination between the different oxidized states when isolating the parent mass within the ion trap. The resolving power of our column was sufficient to clearly separate three species at the parent mass (Figure 4.6a). We extracted the MS/MS data of each LC peak and calculated the extent of identity for the three possible patterns: open, parallel, and cross-over.

For each of the three chromatographic peaks, the primary sequence score (optimized Q depth = 7, 6, 4 for the first, second, and third LC peaks respectively) generated by the target sequence FCGLCVCPCNK was highest among scores generated by the decoy peptides (Figure 4.6c–e), demonstrating a confident match by sequence for the known SNAP25B cysteine-rich region.
Figure 4.6: Identification of the three different disulfide patterns. (a) Extracted ion chromatogram for the mass range of 591.73–591.74 at +/- 8 ppm, where the theoretical monoisotopic m/z for a +2 charge state is 591.748, showing three distinctly resolved peaks. The MS/MS spectra were extracted at retention times 68.03, 71.94, and 77.90 minutes for each respective LC peak. (b) Unique fragment scores. LC elutions of peaks 2 and 3 show the presence of open and parallel patterns (respectively). The dotted line is the 99% confidence line for the presence of a pattern (score = 20, where p-value = 0.01; nd; no detection). (C–e) All three chromatographic peaks have the primary amino acid sequence of the target peptide double disulfide FCGLVCPCNK. We further confirmed our matches by running a decoy type search based upon the input peptide sequence by generating 99,999 sequence variants and plotting the search results for each match. By this metric, we see that matches for the true sequence lie well outside the distribution of false matches. The arrows indicate the primary sequence score for each of the three patterns.
We then calculated the pattern diagnostic score to determine the specific disulfide pattern for each LC peak, with accompanying spectra shown in order of elution (Figure 4.7, Figure 4.8, and Figure 4.9). Peaks 2 and 3 scored above a 99% confidence line for open and parallel patterns, respectively (Figure 4.6b, Figure 4.8, Figure 4.9 and Supplemental Table). By identifying peak 1

Figure 4.7: MS/MS spectra of the precursor mass m/z = 591.73 at retention time 68.03 min. The fragmentation of the peptide with CID shows unique peaks (red spectra = unique fragments for the cross-over pattern, blue = shared fragments amongst three isoforms, and light blue = unique fragments for the open pattern). The numbers next to the cysteine (C) denote the order of position in the primary sequence (FC1GLC2VC3PC4NK). The peaks within the region beneath the bold black line are magnified by x5. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: y2-CO, c9-H2O, z11-CO, b11-NH3, y11-CO, z11, and c10. The details of the peaks are presented in Supplemental Data Table. The Q depth of this analysis was optimized to Q=7. The +SSH and −H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision-induced dissociation (persulfide and dehydroalanine, respectively).
as the SNAP25B cysteine-rich double disulfide peptide, achieving complete chromatographic separation between the three peaks, and confidently assigning LC peaks 2 and 3 as the open and parallel patterns, we conclude that peak 1 is the elution of the cross-over pattern (Figure 4.5 and Figure 4.7). This result is consistent with the observation that the patterns that gave the highest

![Peptide Sequence](image)

**Figure 4.8:** MS/MS spectra of the precursor mass m/z=591.73 at retention time 71.94 min. The fragmentation of the peptide with CID shows significant amount of unique peaks (red spectra = unique fragments for the open pattern, blue = shared fragments amongst the three isoforms, and light blue = unique fragments for the parallel pattern). The peaks within the region beneath the bold black line are magnified by ×5. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: c9-H2O, a5-NH3, a5, z11-CO, b11-NH3, y11-CO, z11, a9, and z10. The details of the peaks are presented in Supplemental Data Table. The Q depth of this analysis was optimized to Q= 6. The +SSH and −H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision-induced dissociation (persulfide and dehydroalanine, respectively).
primary sequence score (refer to Figure 4.6c–e) are consistent with the identification of unique
fragments for the respective patterns (Figure 4.6b).

4.4.6 Triple backbone cleavages
In an ion trap, a cross-linked molecule may break multiple times internally before a breakage
results in separate fragments that then fall out of resonance. The likelihood of multiple
breakages occurring is related to the proportion of amino acids involved in circular paths and
also their fragmentation propensities. After showing that double cleavage events can be a
dominant fragmentation pattern when the disulfide pattern favors it, we reasoned that triple
cleavage events may occur in doubly crosslinked molecules by the same principle of sustained
excitation within an ion trap. The cross-over pattern has a high probability of maintaining its
initial mass after single or double backbone cleavages because of the presence of the two
overlapping disulfide bonds and therefore may have a significant probability of triple backbone
cleavages. We calculated that the cross-over pattern can produce nine unique fragments upon
triple backbone cleavage (Supplemental Table). We calculated the pattern diagnostic score of
the MS/MS spectra of the first chromatographic peak (Figure 4.6a), deductively assigned as the
cross-over pattern, for triple cleavage events. In strong support of our assignment, we detected
two out of nine of these unique fragments 774.36 and 831.60 with a pattern diagnostic score of
17.3, which is above the 95% confidence (where 95% confidence threshold is 13.0; Q= 4; Figure
4.7 and Supplemental Table). In contrast, neither MS/MS spectra of the second or third
chromatographic peaks reached statistical significance for these unique cross-over fragments at
any of the Q depth from 1–10 (Supplemental Table). Consistent with the enhanced
fragmentation at X-Pro bond via CID, both of the observed triple cleavage fragments have a
breakage at the X-Pro bond. Although the thorough validation of triple backbone cleavage by
CID in an ion trap is beyond the scope of this work, this analysis does provide substantial corroborating evidence that we did indeed correctly identify the cross-over peak based on deduction (Figure 4.6b).

4.5 Discussion and conclusions
Here, we report, for the first time, statistically significant identification of all three disulfide patterns for double disulfide species of SNAP25B using CID. Recently, Foley et al. showed that some of the cysteines within the SNAP25B’s cysteine-rich region are oxidized in the rat brain.\textsuperscript{180} They hypothesized that this is through an intradisulfide bond formation. Our data show that this is sterically feasible between all four cysteine residues.

Although the result demonstrates the capacity to identify the three different bond patterns for the initial peptide by using low-intensity peaks, there are clearly limitations inherent with this approach based on the complexity of the problem. First, a mix of the patterns in a single chromatographic peak cannot be resolved to full identification because of the high degree of similarity between each fragmentation spectra – even a small amount of a different coeluting bond pattern can confound identification as the parent masses are inseparable in ion trap fragmentation precursor ion selection. Further, the cross-over variant has no unique fragments after double cleavage and so is only identified by a lack of diagnostic fragments and a higher primary sequence score than the other two variants. Deduction and complete chromatographic separation were sufficient in this case, but a more generalized solution may need to incorporate triple fragmentation cleavage for successful identification.
This study relied on lower resolution ion trap spectra, although MS/MS spectra could potentially have been collected in the Orbitrap at higher resolutions. Two considerations motivated the use of LTQ-generated low-resolution spectra. Primarily, we noticed that unless we sampled near the peak apex, we often did not accumulate enough ions to observe peaks generated from lower frequency breakage events. Furthermore, the spectra we collected away from the apex were

Figure 4.9: MS/MS spectra of the precursor mass m/z = 591.73 at retention time 77.9 min. The fragmentation of the peptide with CID shows significant amount of unique peaks (red spectra = unique fragments for the parallel pattern, blue = shared fragments amongst the three isoforms, and light blue = unique fragments for the open pattern). The peaks within the region beneath the bold black line are magnified by ×5. *Unaccounted peaks in the statistical analysis were checked manually for their identity. From the left: z11-CO, b11-NH3, y11-CO, z11, and c10. The details of the peaks are presented in Supplemental Data Table. The Q depth of this analysis was optimized to Q = 4. The +SSH and H denote the cleaved fragments of the disulfide bond via heterolytic disulfide bond cleavage via collision induced dissociation (persulfide and dehydroalanine, respectively).
typically contaminated with the other disulfide variants that were closely eluting. In particular, intense unique fragments generated by single backbone fragmentation of the open patterns dominated the MS/MS spectra of the other two LC peaks when the separation by LC was insufficient or MS/MS spectra were not taken around the peak apex of the respective LC peaks. This spectral contamination can be observed in spectra from the first and the third peaks of the chromatogram (Figure 4.6a), which both have spectral contamination from the open pattern (661.20 m/z; Figure 4.7 and Figure 4.9). The higher scan speeds of the LTQ ion trap allow sampling near the apex of the chromatographic peaks and thus generate spectra with minimal contamination between the three patterns. The second reason for using spectra derived from the ion trap is that it increases the accessibility of the method to include low-resolution instruments. Future studies will explore the tradeoff between high resolution, spectral contamination, and sampling rate; however, our current statistical analysis and stringent decoy search using low-resolution spectra are clearly adequate to identify and distinguish the patterns under investigation.

Resolving singly bonded patterns on peptides with more than two cysteines is, in some ways, a more challenging endeavor than the double disulfide case. The four cysteines on the SNAP25B cysteine-rich region can form six patterns of single disulfide species. Still, the general approach we have outlined here for dealing with double cross-links should be applicable to the single link patterns and is currently in progress. Perhaps the most significant advantage of the approach used in this work is that it is generally compatible with standard ion trap-based shotgun proteomic analysis. Because ions stay in resonance until they fragment into distinct molecules, only typical ion trap MS/MS CID is required to achieve the double and even triple fragmentation
events necessary to assign identities. It should be possible to review ion trap data sets where no reduction and alkylation were performed and to extract new identities and linking patterns. It should also be possible to add disulfide bonds to the list of searchable modifications with relatively unobtrusive alterations to standard search algorithms. However, special consideration must be given to searching for interpeptide cross-links (whether derived from interprotein or intraprotein events) because of the combinatorial expansion of the search space.

An additional challenge in examining disulfide bond patterns routinely is that disulfide bonds may undergo an interchange reaction with neighboring cysteines at alkaline or neutral pH. The primary purpose of this study was to show that the various disulfide bond patterns of the cysteine-rich region in SNAP25B could be identified, so no action was taken to avoid disulfide interchange during sample preparation. Future studies seeking to examine the relative ratios of disulfide patterns will need to be undertaken at lower pH (2–6), although this must be balanced with concern for diminishing tryptic activity.[26] Sample handling at a pH of 6.0 may be an optimal solution and indeed has been shown to minimize disulfide shuffling while allowing tryptic activity.[14]

4.6 Acknowledgements
We thank Tomonori Baba for the initial development of the program double_disulfide_searcher and data analysis. We would like to thank Aman Makaju and Andrew Mathis for assistance in running MS samples. We thank Nathan Doyle for SNAP25B protein production and MS sample preparation. JTP and RMT were supported from generous institutional start-up funds.
Chapter 5  Automatic lipid classification by machine learning

Author’s Note: This chapter describes an automatic classification tool, trained on the LIPID MAPS ontology to provide simple classification for any lipid structure. I wrote the algorithms and the paper. These results are under review in Bioinformatics.  

5.1 Abstract

Motivation: Modern lipidomics is largely dependent upon ontologies because of the great diversity exhibited in the lipidome; no automated lipid classification exists to facilitate this partitioning. The size of putative lipidome far exceeds those currently classified, despite a decade of work. Ongoing classification of unknown lipids will require automated tools to reduce the time requirement, improve the resulting accuracy, assist in manual classification workflows, and facilitate mass spectral analysis.

Results: We introduce a 99% accurate automated classification tool of any given lipid into the LIPID MAPS ontology as well as all feature sets required for alternative implementations. The classification is trained by machine learning upon simple chemical characteristics. The decision trees produced are intelligible and can be used to clarify implicit assumptions about the current LIPID MAPS classification scheme. By applying the classifier to the LIPID MAPS database, we also discovered many hundreds of lipids that are currently misclassified, strongly underscoring the need for an automated classifier like this.

Availability: Source code and chemical structure lists as SMARTS search strings are available under an open-source license at https://www.github.com/princelab/lipid_classifier

---

4 Publication authors are Taylor RM, Miller RM, Miller RD, Porter M, Dalgleish J, Prince JT
Supplemental information: Supplementary data are available at *Bioinformatics* online.

5.2 Introduction

Lipids are a fundamental component of biological systems and perform diverse roles in many cellular pathways. They comprise several thousands of structurally distinct species whose diversity is preserved by dedicated cellular systems. The lipid composition of a cell is linked to its function, hence lipids are excellent subjects for gaining insight into biological systems and predicting abnormalities. Indeed, lipids are known to play a major role in diverse diseases afflicting millions, including obesity, diabetes, asthma, hypertension, arthritis, and cancers. Lipidomics—the analysis of the lipid composition, localization, and activity of a cellular or physiological system—is a burgeoning field of research.

One major difficulty in working with lipids is dealing with their great structural diversity. To help address this challenge, the LIPID MAPS Consortium has created and been refining the LIPID MAPS Lipid Classification System (LMLCS) which has become the de facto ontology used in lipid research. With this ontology, the lipid research community is able to discuss predicted lipid properties and cellular functions in ways that would otherwise be impossible.

Indeed, classification of biomolecules is a prerequisite to any systems biology approach. This is particularly true in the area of mass spectrometry identification, where theoretical fragmentation spectra (used for matching with each experimental spectrum) are generated by different sets of rules based on a biomolecule's type (protein, metabolite, etc.). The principle holds true for lipids: nearly all existing identification approaches require that a lipid be classified (although sometimes implicitly) in order to generate a theoretical spectrum from a lipid's
structure. Classification makes possible restricted search space structural comparisons and even such fundamental tasks as representing lipids in a systematic fashion.

The benefits resulting from classification are definite; these benefits are currently inaccessible to lipids which have not been previously classified. Over the last ~15 years, LIPID MAPS has classified over 38,000 lipids, an enormous feat. Still, there are likely in excess of 120,000 lipid species and probably more when considering oxidative modifications, yet undiscovered natural products, and unanticipated future synthetic modifications. And, although automatic classification tools have been alluded to, currently there is no publicly available software for the automated classification of lipids. Although classification can be performed manually, manual classification cannot be used in any automated software pipelines, does not scale well, and may not always be accurate (Danziger et al., 2011).

We present an approach to generate a classifier trained on the LIPID MAPS ontology and structural database (LMSD) which can be used to classify novel lipids automatically and assist in manual classification workflows.

5.3 Methods
More extensive methodology is given in the methods supplement.

5.3.1 Chemical Language and Identifying features
We used Rubabel, a cheminformatics software suite built upon the OpenBabel library, to provide programmatic representation of chemical structures which were searched by SMILES Arbitrary Target Specification, or SMARTS, search strings to produce a list of chemically identifying structural characteristics, as detailed in Supplemental Table 1. Each identifying
structural characteristic formed a binary variable indicating presence of a feature or representing a numerical count of feature matches.

5.3.2 Classification by Machine Learning

WEKA machine learning algorithms were compared by several numerical performance measures (see Algorithm Selection supplemental). The J48 decision tree algorithm was selected based on performance, speed, and interpretability. Classification accuracy was optimized upon a 15% subset of the LMSD. The WEKA-produced decision trees contain a rule-by-rule determination of lipid classification based on the identifying structural characteristics and were trained upon a randomized 90% sampling of the entire training dataset.

Each lipid of the training dataset, which consisted of the entire LMSD, was structurally analyzed to produce a structure feature list which was then split into the hierarchal levels of the LMSD ontology (see Figure 5.1, panel A). Each feature list was then analyzed by WEKA to produce decision trees representing the classification steps at every hierarchical level.

This generated hierarchy of classification trees are dynamically loaded into a programmatic classification system (see Figure 5.1, panel B) implemented in Ruby, which provides a classification for a given lipid structure, by 1) generation of a structural feature list, and 2) application of each hierarchical decision tree.

The Ruby classification system was evaluated for accuracy across all hierarchical levels by examination of the entire LMSD. Each classification was considered a miss or hit in two categories, category classification and category-internal classification. These two scores present scores for both the complete classification and the relative importance of category selection.
Classification model accuracy was confirmed by manual evaluation of all misclassified lipids and annotated any potential ontology changes. When the classification models assigned lipids to categories that were not indicated by their structure, we identified relevant structural features.

Figure 5.1: Lipid classification workflow schematic. Panel A) A lipid classification is constructed from a training set of lipids with known classifications. Each lipid in the training set is analyzed structurally to produce a list of structural characteristics which can be used for machine learning analysis. These feature sets are split into hierarchical groups according to their classification. Machine learning analysis builds a distributed set of decision trees at each hierarchical level. Panel B) A novel lipid is analyzed structurally, then the structural analysis is analyzed by consultation with the WEKA produced decision trees at each hierarchical level to generate a complete classification. Overlaid boxes highlight the contributions of each hierarchical layer to a complete classification.
which could provide correct classification assignment.

5.3.3 Evaluation of novel lipids
Evaluation of the trained classifier was performed upon an extracted subset of the LipidBank database, which consisted of some 1195 molecules, many of which are similar to molecules which are contained in the LMSD. Accuracy was measured by manual evaluation to determine if these lipids were 1) properly categorized into, and 2) fit within the LMSD ontology. Hits and misses were only counted when the lipid could be classified into the LMSD.

5.3.4 Statistical considerations and model validation
Algorithm selection was performed by split-percentage validation at 66%. WEKA derived accuracy scores for these data were compared with the J48 model based on 90% of the training dataset with cross-validation.

5.4 Results

Figure 5.2: Representative decision trees for LMSD classifications. A) Glycerolipid category (GL) into 6 class levels, GL00-GL05, based upon chemical features. B) Fatty acyl (FA) category demonstrates increasing complexity.

5.4.1 Classifier and feature selection
The structural characteristics and machine learning algorithms chosen ensure the classification model decision trees are human understandable (see Figure 5.2). The resulting comprehensive
classification provides an automatic classification that retains human interpretability. Analysis of a lipid can be performed in less than a second, making automated, and on-the-fly classification possible.

5.4.2 Classifier model performance
The comprehensive reference implementation in the Ruby language provides less than 1.2% error across all classification levels (see Misclassifieds supplemental). An equivalent implementation trained on a 66% split percentage yields 3.0% error, and suggests that the chemical features selected are largely robust attributes for classification. At the category level, we reach 99.98% accuracy (as described in Table 5.1) when suggested improvements to the existing ontology are followed as outlined in the supplemental tables 2 and 3.

Table 5.1: Classifier performance for entire LMSD and categories slices of the LMSD. Category level errors represent misclassifications which put a lipid into the wrong category and are excluded from ‘within category’ error counts. Within category errors represent any misclassification of a lipid other than a category level error.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of lipids</th>
<th>Category Level Error counts (%)</th>
<th>Within Category Error counts (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire LMSD</td>
<td>36785</td>
<td>3 (0.01%)</td>
<td>426 (1.16%)</td>
</tr>
<tr>
<td>Fatty Acyl [FA]</td>
<td>5763</td>
<td>1 (0.02%)</td>
<td>3 (0.05%)</td>
</tr>
<tr>
<td>Glycerolipids [GL]</td>
<td>7538</td>
<td>0 (0.00%)</td>
<td>1 (0.01%)</td>
</tr>
<tr>
<td>Sterol Lipids [ST]</td>
<td>2561</td>
<td>0 (0.00%)</td>
<td>16 (0.62%)</td>
</tr>
<tr>
<td>Prenol Lipids [PR]</td>
<td>1193</td>
<td>1 (0.08%)</td>
<td>0 (0.00%)</td>
</tr>
<tr>
<td>Sphingolipids [SL]</td>
<td>1293</td>
<td>0 (0.00%)</td>
<td>0 (0.00%)</td>
</tr>
<tr>
<td>Polyketides [PK]</td>
<td>6744</td>
<td>0 (0.00%)</td>
<td>11 (0.16%)</td>
</tr>
<tr>
<td>Sphingolipids [SP]</td>
<td>3934</td>
<td>0 (0.00%)</td>
<td>385 (9.79%)</td>
</tr>
<tr>
<td>Glycerophospholipids [GP]</td>
<td>7759</td>
<td>1 (0.01%)</td>
<td>10 (0.13%)</td>
</tr>
</tbody>
</table>

5.4.3 Novel lipid analysis
Evaluation of classified lipids from the LipidBank database demonstrates the capability of this classification to handle novel lipid classes. In 656 novel lipids, we correctly classified 70% of the lipids which were manually determined to be within the current LIPID MAPS ontology (see Novel Lipid Analysis supplement).
5.4.4 Ontology modifications
We carefully inspected the results of the model built upon cross-validation training and discovered ~150 lipids that are misclassified in LIPID MAPS (as detailed in the Misclassified supplemental). The number of misclassified lipids indicates the difficulty of hand-curating a database the size and complexity of LIPID MAPS. In addition, the ability of our classifier to find these misclassified lipids strongly supports the utility of our automated approach.

5.5 Discussion
5.5.1 Classifier design and performance
Despite a number of misclassified lipids, the high degree of accuracy we achieved suggests that the LIPID MAPS ontology is structurally coherent. Additionally, despite an origin in synthetic pathways, the 2009 revision of the ontology is largely structurally derived, making an analysis like this successful. The chemical characteristics chosen provided robust machine learning attributes across a majority of potential classifier algorithms (see the Algorithm Selection supplement). The J48 algorithm was chosen because it provided nearly identical performance to the highly performant LMT algorithm, but at 1/200\textsuperscript{th} of the analysis time.

While the classifier performance is largely acceptable, the neutral glycolipid class of the sphingolipid [SP] category remains a source of error. This is due to several factors, one of which is the degree to which the ontology fails to encompass the diversity of the glycolipids. There are ~3000 lipids within this class and many classifications of these lipids are dependent upon sugar oligomer length to differentiate nomenclature precedence. Thus, resolution of this issue will require 1) consideration of ontology changes to better reflect the diversity of neutral glycolipid structures, and 2) consideration of structural characteristics beyond the SMARTS systems currently employed, such as a longest-path finding algorithm. The current diversity of
glycolipids within the LMSD fails to encompass possible diversity, as there are only a few sugar monomers contained therein, fucose, mannose, galactose, and glucose. We suggest that future efforts investigate this diversity more fully and propose more sweeping ontology changes, such as subclasses which allow for classification of unrepresented sugar monomers, and which only contain lipids with a specified sugar root structure.

5.5.2 Novel lipid analysis
Evaluation of a novel lipid library demonstrates the capacity of our classifier to streamline novel lipid analysis. It still further demonstrates need for further refinements within the established ontology, as many of these lipids would be very appropriately grouped into ontology which is not yet represented within the LMSD

5.5.3 Ontology modifications
Many of the aforementioned misclassifications are due to small structural differences. Lipid LMGP04040006 is classified as a dialkylglycerophosphoglycerol. Examination of the structure demonstrates an acyl group in place of an alkyl group, corresponding to our classifier's assignment for this lipid as a 1-acyl, 2-alkylglycerophosphoglycerol, or LMGP0411. The fatty acid LMFA01010053, in the straight chain fatty acid class, is clearly branched, corresponding to the reclassification suggested by our analysis.

The model excelled at assigning lipids that contain multiple structural features. Several fatty acids are both branched and unsaturated. These fatty acids are distributed among both the unsaturated fatty acids and the branched fatty acids even though they are structurally similar. The model followed the established ontology that branching takes precedence over unsaturation and assigned these lipids to the correct classification.
In accordance with IUPAC guidelines\textsuperscript{205–210}, neutral glycosphingolipids were assigned a group based on their root sugar chain, the first four sugars and their linkage to the sphingolipid. LIPID MAPS suggested nine groups (or series, LMSP0501-09). There are several sets of distinct lipids within the Neolacto subgroup that do not fit into it nor any other group. These sets (LMSP0505DC-F and LMSP0505DM-N) contain 32 and 16 lipids with two unique roots. We suggest implementation of two new ontology groups for these distinct roots: gluco-globo (LMSP0510) and galacto-lacto (LMSP0511). Gluco-globo highlights the similarity to the isoglobo (LMSP0506) series, excepting the terminal N-acetyl glucosamine. Galacto-gluco illustrates the relationship to the Gala series (LMSP0509) as it contains repeated galactose monomers and highlights the terminal glucose monomer. These new ontologies would represent the incorrectly classified lipids in their own ontology and are suggested to improve classification of current and future neutral glycosphingolipids.

5.5.4 Future Directions
Future work should expand the classification system to classify non-lipids into general categories, and improve upon some existing limitations of the extensive sugar nomenclature within the sphingolipid category. Future efforts will shorten analysis time per lipid. Future work should evaluate whether the need for an alternative ontology which enables multiple classifications for a given lipid exists, such as the aforementioned branching and unsaturation precedence. An ontology capable of multiple labels might more accurately reflect the diversity of the lipidome.
Chapter 6  Msplinter: a structure driven approach to lipid fragmentation prediction

Author’s Note: This chapter introduces a fragmentation prediction algorithm for lipids which is capable of predicting the complex cyclization products published in lipid fragmentation studies. I present the implementation and application to experimental spectra and biological shotgun lipidomic analysis. 

6.1 Abstract

Motivation: Current lipid fragmentation algorithms do not predict the products of complex fragmentation mechanisms which are well characterized in published literature. Many identifications cannot be made from MS1 data alone, requiring MS/MS based identification algorithms. Current fragmentation models do not predict complex fragmentation events as evidenced in published literature. A chemically derived fragmentation model can provide chemically sound fragmentation products which more accurately reflect the reality of lipid fragmentation pathways as determined by decades of fragmentation research.

Results: We introduce a chemically derived fragmentation model which predicts lipid fragmentation products by electron-pushing mechanisms implemented in code. The predicted masses are accurate and provide statistically relevant discrimination of ceramide lipids from among theoretical and experimental datasets.

Availability: All relevant source code can be found at https://www.github.com/princelab/msplinter.

________________________

5 Publication authors are Taylor RM, Miller RH, Anthonymuthu TS, and Prince JT
6.2 Introduction

Lipids are fundamental components of biological systems and perform diverse roles in vital cellular pathways. The lipid composition of a cell is highly predictive of cellular system abnormalities\(^{182}\) and lipid abnormalities play causative or contributory roles in diverse diseases.\(^{183–193}\) Lipidomics—the analysis of the lipid composition, localization, and activity of a cellular or physiological system— is a field of rapidly increasing importance.

Mass spectrometry (MS) provides the throughput, sensitivity, and concurrency required to quantify and identify lipids; however, converting MS signal into lipid identities is highly challenging. The bottleneck in mass spectrometry based lipidomics is still data analysis,\(^{16,22}\) and arguably the most difficult step lies in identifying lipid species from fragmentation spectra. Several lipids generate characteristic reporter ions or neutral losses. While these patterns are helpful in identifying particular head groups, they are only a first step towards obtaining

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Coverage</th>
<th>No Training Set Requirement</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>On-the-fly</th>
<th>Flexibility</th>
<th>Abundances</th>
<th>Complex fragment prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental library</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>=</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Characteristic ion fragment tables</td>
<td>=</td>
<td>=</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>=</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Bond cleavage models</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+/−</td>
<td>−</td>
</tr>
<tr>
<td>CDF(proposed)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>=</td>
<td>+</td>
<td>=</td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Table 6.1: Comparison of approaches for MS/MS lipid fragmentation prediction algorithms.

confident identifications by MS/MS. For instance, isobaric and/or closely eluting lipids often produce chimeric spectra upon fragmentation. Disentangling the fragment ion contributions
from 2 or more spectra are highly challenging without predictions of the complex and unique product ions characterized in literature.

6.2.1 Current prediction algorithms are deficient

Current lipid fragmentation approaches are deficient in one or more ways: they provide identifications only for certain lipids, lack the capability to predict complex fragmentation events, and/or require a significant investment in generating a new experimental training set in order to analyze new lipids. Table 6-1 summarizes the relative strengths and weaknesses of common approaches.

Current lipid fragmentation approaches can be divided into three primary categories, 1) experimental library\textsuperscript{18,211} lookups, 2) tables of characteristic fragments,\textsuperscript{11,21,212–217} and 3) bond cleavage models.\textsuperscript{19,20}

Libraries of experimentally collected fragmentation spectra represent complete truth for lipid identification but the requirement for spectral collection can be especially prohibitive when purified lipids are not available. Experimental spectra must typically be collected on similar instrumentation with similar experimental parameters for comparative identification methods to be successful. Experimentally derived standard spectra are valuable contributions but they cannot be applied to previously uncharacterized lipids without considerable effort.

Tables which predict characteristic fragments are easily implemented because they consider only the easiest and most predictable fragmentation events by lipid group. Bond cleavage models consider all possible bond cleavages and are subsequently scored by energy to determine the most likely fragmentations. However, neither approach predicts the complex fragmentation schemes which are well characterized in primary literature.\textsuperscript{218–240}

107
Current lipid fragmentation algorithms fail to capture the rich and diverse fragmentation pathways characterized by decades of research into lipid and metabolite fragmentation pathways. These data are neglected or implemented only indirectly by existing software tools. As they stand, current implementations cannot perform the low level chemical manipulations required to fully detail complex fragmentation schemes such as the cyclizations observed in published literature.

6.2.2 Chemically derived fragmentation model
To address the limitations of existing models, we introduce a chemically derived fragmentation (CDF) tool which models, \textit{in silico}, the behavior of fragmentation as ‘electron pushing’ schemes familiar to chemists. Each scheme is derived directly from literature studies of fragmentation pathways.

Lipids from different lipid classes fragment differently from one another\textsuperscript{241,242} and a well-established practice in fragmentation software treats lipids on a class-by-class basis, or in a 'divide-and-conquer' approach.\textsuperscript{196–199} We have adopted this approach as well, as the complex fragmentation schemes are highly specific to specific lipid classes. We evaluate the CDF approach upon lithiated ceramide lipid samples from literature and experimental sources as characterized by Hsu et al. in 2002.

6.3 Methods
6.3.1 Model each lipid chemically
Each analyte molecule is programmatically loaded as a chemical model, in any of the 125 different chemical file formats supported by the OpenBabel C++ chemical toolkit.\textsuperscript{244} Each chemical model is manipulated using Rubabel,\textsuperscript{201} a framework built on top of OpenBabel, which we recently developed in order to rapidly and easily implement fragmentation events.
6.3.2 Fragmentation scheme implementation

Each fragmentation pathway was implemented as chemical structure test cases, according to principles of behavior driven development. Once a test case was implemented, the mechanism of the reaction was encoded in chemically derived steps which act upon the chemical model, as shown in Figure 6.1, and refined until all test cases were passed.

6.3.3 Search model for known fragmentation motifs

Each chemical model is searched by SMARTS search string for all known fragmentation motifs and appropriate fragmentation methods are selected to be applied to the molecule. Each model is processed to set the experimental adducts, ion mode, and fragmentation mode. These parameters are set so fragmentation schemes can be selected as compatible with the experimental conditions.
6.3.4 Application of fragmentation schema
Each reaction occurs as a combination of bond formation, cleavage, and movement of electrons within the chemical model, precisely as the published fragmentation schemes indicate. The products of the algorithm are themselves chemical models. Rules are signified as primary fragmentation pathways, as secondary rearrangement pathways, or both, which governs how the fragments are further processed.

6.3.5 Comparison to characterized product ions
The performance of the rule sets was initially evaluated by correlation of predicted product ion masses with the published ion masses.

6.3.6 Spectral scoring algorithms
In order to assess spectral matching quantitatively, we implemented several spectral scoring algorithms: F1 score, spectral hit count, and the probabilistic score from the proteomics tool Andromeda.\(^ {23} \)

F1 score is the harmonic mean between recall and precision, and provides a robust statistical measure of absolute Type I and Type II errors. Recall and precision are defined in terms of true positives (TP), false positives (FP) and false negatives (FN). Recall is defined as \( TP / (TP + FN) \) and precision is defined as \( TP / (TP + FP) \).

\[
F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

The spectral hit count is a simple weighted measure of hits, misses and unpredicted ions between the two spectra. X, Y, and Z represent scaling factors which were manually optimized to produce an approximate ‘no-match’ threshold value near 0 for preliminary spectra, making X=10, Y=1, and Z=2.
The Andromeda score is derived from the probability of matching a certain peak, and is optimized by $q$, or the number of peaks in a given $\sim100 \text{ m/z}$ bin which are considered in the analysis. The formula is given here,

$$s(q, loss) = -10 \log_{10} \sum_{j=k}^{n} \left[ \left( \frac{n!}{k! (n-k)!} \right) \left( \frac{q}{100} \right)^j (1 - \frac{q}{100})^{n-j} \right]$$

where $n$ is the total number of theoretical ions, and $k$ is the number of matching ions in the spectrum. As this fragmentation model has not been implemented before, each score was compared to determine the best score for lipidomic comparison upon the experimental dataset. Preliminary work suggested that the F1 and HitCount scores are more robust at differentiating spectral matches. All scores were generated and reported at an optimized $q$ depth, analogous to the Andromeda algorithm scores but considered without the $100 \text{ m/z}$ bin.

### 6.3.7 Experimental Spectra

Ceramide class lipid standards from Avanti Polar Lipids (Alabaster, AL) were suspended in Folch reagent (2:1 chloroform:methanol)\textsuperscript{247} with 15 mM LiCl as the adduct species, except for the ceramide phosphate sample which required 30 mM LiCl. Samples were injected at 8 $\mu$L/min using a direct-inject Thermo Scientific (Waltham, MA) IonMax electrospray ionization (ESI) soft-ionization spray head from a Hamilton (Reno, NV) GASTIGHT glass syringe. The spray voltage and capillary were maintained at 5.0 kV and 275 C with an arbitrary sheath gas ($\text{N}_2$) flow rate of 8. The selected 7 ceramide class lipids are listed in Table 2. All MS/MS and MS3 fragmentation spectra were collected at 100k resolution in the FT mass analyzer of an Orbitrap XL (Thermo Scientific) mass spectrometer. Spectra were post-processed into a normalized cumulative spectrum containing both MS2 and MS3 peaks by Mspire.\textsuperscript{177,248}
We validated algorithm accuracy by conducting comparison searches upon each experimental spectra against all lipids within $\pm$ 5 amu from the parent mass as selected from the LIPID MAPS structural database (LMSD). Each lipid was identified by consensus spectral scores by highest score. For the F1 and spectral hit count scores, statistical relevance was estimated by sigma deviations from the mean, where anything greater than 3$\sigma$ was considered significant.

6.3.8 Complex samples preparation

For isolation of lipids, murine lung cell pellets were re-suspended in 900 $\mu$l ice-cold chloroform/methanol (1:2) and incubated for 15 minutes on ice, then briefly vortexed. Separation of aqueous and organic phases required addition of 400 $\mu$l of ice-cold water and 300 $\mu$l of ice-cold chloroform. The organic phase was collected into a fresh vial, and lipids were dried in a vacuum centrifuge (Eppendorf Concentrator Plus). Lipids were characterized and quantified using a shotgun lipidomics technique on a Thermo Scientific LTQ Orbitrap XL mass spectrometer. Evaporated lipid samples were re-suspended in a 2:1 chloroform:methanol Folch solution (200 $\mu$L). The re-suspended lipids were then combined with a Folch solution with 15 mM LiCl acting as an ionizing adduct species. A cocktail of internal standards was spiked into each sample for mass calibration and characterization data alignment. Samples were analyzed using a 10 minute mass-window scanning method in positive-ion mode at a resolution of 100,000 (FWHM at 400 m/z) for all primary MS1 scans. MS2 fragmentation data was also collected and manually verified to give additional confidence to the correct identification of abundant lipid species. Samples were injected at 8 $\mu$L/min using a direct-inject Thermo Scientific IonMax electrospray ionization (ESI) soft-ionization spray head from a Hamilton GASTIGHT glass syringe. The spray voltage and capillary temperature were maintained at 4.5 kV and 275° C respectively. Fragmentation events were performed on the most abundant
peak from each survey scan (150-2000 m/z) with dynamic exclusion settings of 500 items, which expire after 600 seconds.

6.3.9 Shotgun lipidomic analysis

We produced a reference identification and quantitation analysis by our standard workflow as follows. Data were analyzed using an in-house developed peak summarization, recalibration, and MS1 lipid identification software using lipid database information from the LIPID Metabolites and Pathways Strategy (LIPID MAPS) Lipidomics Gateway database. To ensure high-confidence identifications, an intensity threshold estimated to be 5% above instrumental static signal was implemented. The lipid species identified across different ionization states or with adducts were totaled together. Quantification was completed by normalizing total ion counts to the relative abundance of the internal standard that was spiked into each sample.

We analyzed the shotgun lipidomic data set by comparing them against CDF generated spectra. Potential lipid matches were extracted from the LIPID MAPS database in a +/- 50 ppm window. For each potential match, fragmentation masses were predicted by CDF and compared against the experimental fragmentation spectra by the above scores. Statistical significance was again estimated by sigma level.

6.4 Results

We successfully captured the schemes reported by Hsu et al. in our chemically derived fragmentation model. The coded implementation retains chemical relevance by using chemical terminology. Each rule is itself modular and independent and can be selected as needed. We tested each fragmentation rule on chemical test structures derived from published schemes individually to ensure that each rule produces the expected fragment ions.
Table 6.2: Collision induced dissociation ions characteristic of selected Ceramides from Hsu et al., 2002. Green shading indicate cyclized product ions. Additional fragmentation comparison data is found in the supplemental materials. \(^d\) indicates a diagnostic ion which is unique to that lipid or fragmentation pathway, as characterized in the original publication.

<table>
<thead>
<tr>
<th>LMID</th>
<th>common ions (m/z)</th>
<th>Fatty acyl ions (m/z)</th>
<th>Long chain base ions (m/z)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a1</td>
<td>a2 (a1-30)</td>
<td>a1-18</td>
</tr>
<tr>
<td>LMSP020100012</td>
<td>638</td>
<td>608</td>
<td>620</td>
</tr>
<tr>
<td>LMSP02010009</td>
<td>636</td>
<td>606</td>
<td>618</td>
</tr>
<tr>
<td>LMSP02010004</td>
<td>526</td>
<td>496</td>
<td>508</td>
</tr>
<tr>
<td>LMSP02010006</td>
<td>554</td>
<td>524</td>
<td>536</td>
</tr>
<tr>
<td>LMSP02020012</td>
<td>640</td>
<td>610</td>
<td>622</td>
</tr>
<tr>
<td>LMSP02020008</td>
<td>556</td>
<td>526</td>
<td>538</td>
</tr>
<tr>
<td>LMSP02020029</td>
<td>544</td>
<td>514</td>
<td>526</td>
</tr>
<tr>
<td>LMSP02020030</td>
<td>572</td>
<td>542</td>
<td>554</td>
</tr>
<tr>
<td>LMSP02030015</td>
<td>560</td>
<td>530</td>
<td>542</td>
</tr>
<tr>
<td>LMSP02030004</td>
<td>656</td>
<td>626</td>
<td>638</td>
</tr>
<tr>
<td>LMSP02030005</td>
<td>684</td>
<td>654</td>
<td>666</td>
</tr>
</tbody>
</table>
We tested all the fragmentation rules on a consensus set of 11 lipids found in both the Hsu et al., 2002 paper and the LMSD. We compared published fragment masses for these 11 lipids against our CDF tool in Table 6.2. For these 11 lipids, we successfully predict all of the major fragment masses.

We compared our lipids to 7 experimental spectra (as listed in Table 6.3 and the Spectral_comparison supplement) against the predicted fragments of isobaric species. This analysis was performed on high mass accuracy data and requires greater accuracy in chemical composition to match than the comparison to the published masses. The CDF scores identify each the candidate lipid from all other isobaric species in the LMSD selection.

Table 6.3: Msplinter prediction scores of 7 experimental spectra against isobaric species +/- 5 amu from the experimental peak. The identity is assigned by highest scores in the HitCount metric. F1 scores largely tracked HitCount scores and can be found in the Spectral_comparison supplement. The C6 ceramide is not represented in the LMSD as is instead referenced by SMILES and file name.

<table>
<thead>
<tr>
<th>LMID or SMILES</th>
<th>Name</th>
<th>Predicted identity</th>
<th>Hit Count</th>
<th>σ Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMSP02010014</td>
<td>C2 Ceramide</td>
<td>LMSP02010014</td>
<td>104</td>
<td>19.82105653</td>
</tr>
<tr>
<td>CCCCCCCCCCCCCC/C= C/[C@@H][[C@@H](C O)NC(=O)CCCCC)O</td>
<td>C6 Ceramide</td>
<td>C6.mzML</td>
<td>77</td>
<td>17.4396082</td>
</tr>
<tr>
<td>LMSP02010002</td>
<td>C12 Ceramide</td>
<td>LMSP02010002</td>
<td>81</td>
<td>12.36655111</td>
</tr>
<tr>
<td>LMSP02010001</td>
<td>C14 Ceramide</td>
<td>LMSP02010001</td>
<td>17</td>
<td>7.204528068</td>
</tr>
<tr>
<td>LMSP02010020</td>
<td>C17 Ceramide</td>
<td>LMSP02010020</td>
<td>107</td>
<td>9.364938564</td>
</tr>
<tr>
<td>LMSP02020012</td>
<td>C24 Sphinganine</td>
<td>LMSP02020012</td>
<td>133</td>
<td>11.71072543</td>
</tr>
<tr>
<td>LMSP02050002</td>
<td>N-Palmitoyl-Ceramide Phosphate</td>
<td>LMSP02050002</td>
<td>72</td>
<td>10.18440042</td>
</tr>
<tr>
<td>LMSP02030001</td>
<td>N-Stearoyl-phytosphingosine</td>
<td>LMPK12020069</td>
<td>10</td>
<td>4.182847713</td>
</tr>
</tbody>
</table>
Shotgun lipidomic identifications are presented in the Shotgun supplementary tables. Statistical analysis of the CDF results demonstrate 4 distinct species which were significant above a 3σ level, out of the 21 unique fragmentation spectra collected. The MS1 based reference analysis identified over 300 lipids. The majority these lipid identifications are unique, with only one specific lipid being confidently identified in both samples.

6.5 Discussion
We have introduced and demonstrated the capability of a CDF based fragmentation model to predict complex fragmentation products. Each product is a molecule itself and this recursive algorithm structure allows for prediction of fragment ion products which only exist as the stable end-products of chemical fragmentation events.

Our three scoring methods demonstrated varying abilities to discriminate between spectral matches. We found that the Andromeda algorithm was unable to differentiate trues and falses from our CDF fragmentation algorithm, as seen in the supplemental tables, which we attribute to
the small number of matched fragment ions. This discrepancy is a product of the limited fragmentation efficiency of MS2 vs an optimized multi-level MS(n) fragmentation protocol. The F1 score didn’t perform as consistently as the HitCount metric which we attribute to a dependency upon false positive estimation from match, miss, and unpredicted counts.

6.5.1 Importance of modeling cyclizations
The importance of cyclization events in predicting ion spectra is shown in Figure 2 by highlighting the contributions to a fragmentation spectra which arise from the complex fragmentation event. Major fragmentation products are products of cyclization reactions which provide diagnostic ions for identification purposes (see Table 1). Without consideration of these cyclizations, unambiguous identification is often not possible.

6.5.2 Published spectra comparison
CDF models are able to predict the majority of fragment species as elucidated by Hsu et al. No existing lipid fragmentation software claims to model the predicted cyclized fragmentation products. Among these 11 lipids, we demonstrate consistent performance which demonstrates the reliability of a chemically derived model.

6.5.3 Experimental spectra comparison
When used to match theoretical spectra against experimental spectra, the Msplinter CDF model is capable of discriminating the true lipid from among 200-400 isobaric lipid species in nearly all cases. The comparison is biased towards ceramides by the limitations of the currently available fragmentation rules which only characterize ceramide fragmentations. The phytosphingosine standard exemplifies this, as it is poorly predicted because of the significant structural differences which eliminate many of the ceramide fragmentation pathways. A phytosphingosine example was intended to demonstrate the capabilities of this CDF model to predict specific
fragmentation events. Implementation of fragmentation schemes appropriate to phytosphingosine appear to be necessary to successfully identify the lipid by MS2 fragmentation alone.

For our experiments matching experimental spectra we used a broad, 10 m/z window. Thus, there were a considerable number of compounds that had opportunity to match with the spectra (200-400, see supplemental tables). Coupling an MS/MS database search with a more realistic precursor tolerance (e.g., 5ppm) would, of course, yield additional discrimination over what was demonstrated here.

6.5.4 Shotgun Lipidomic Comparisons
Preliminary comparisons between the reference shotgun lipidomic analysis and CDF derived analysis demonstrates the complimentary nature of a MS1 and MS2 analysis. We confirmed the presence of 2 lipids which were identified in both analyses and which both identifications agreed, despite this lipid being a glycerophospholipid. The limited applicability of ceramide rules to other lipid classes highlights the strengths of our modular approach. The rest of the CDF identifications demonstrate much higher scores on ceramide samples which demonstrates the capability of CDF models to selectively and specifically predict the fragmentation of a specific class of lipid.

In comparison to phytosphingosine and to general lipid extractions, we observe that without specific rules, fragmentation spectra cannot always be adequately predicted. Our CDF implementation rules are entirely modular and can be selected as needed to optimize for specific analysis needs, or compared against standard spectra to ascertain which rules can predict the observed masses. CDF can therefore be used as chemically grounded building blocks for a
machine learning derived selection of fragmentation rules for previously uncharacterized fragmentation conditions.

6.6 Conclusions
A decade of work in lipid fragmentation pathway characterization is available as a basis for chemically derived fragmentation models. We present replication of experimentally observed fragment ions including highly diagnostic cyclization product ions. We demonstrate that we can correctly identify ceramides from among isobars, even without high mass accuracy precursor mass selection. This work is complimented by a recently developed automatic lipid classification tool based on machine learning and analysis of novel lipids to determine which are amenable to these ceramide fragmentation schemes.\textsuperscript{250}

In future work, we will consider the feasibility of a more flexible programmatic reaction scheme, such as SMIRKS (Daylight CIS). We will also implement additional fragmentation prediction schemes to add fragmentation predictions for additional lipid classes. Further efforts will examine machine learning as a method for optimizing rule selection to increase specificity of current predictions, better match fragmentation conditions, and to predict fragmentation of previously uncharacterized lipids.

The basic framework now exists to implement additional sets of fragmentation prediction reactions based upon other published works which target alternative experimental conditions, sample preparations, and lipid classes.
Chapter 7  
Massifquant: open-source Kalman filter based XC-MS isotope trace feature detection

Author’s Note: This chapter represents an open source implementation of a Kalman filter capable of detection of isotope traces from MS survey scan data. I contributed to the distillation of ideas and preparation of the manuscript. These results were published in Bioinformatics.  

7.1 Abstract

Motivation: Isotope trace detection is a fundamental step for XCMS data-analysis that faces a multitude of technical challenges on complex samples. The Kalman filter application to isotope trace detection addresses some of these challenges; it discriminates closely eluting isotope traces in the m/z dimension, flexibly handles heteroscedastic m/z variances and does not bin the m/z axis. Yet the behavior of this Kalman filter application has not been fully characterized since no cost-free open-source implementation exists and incomplete evaluation standards for isotope trace detection persist.

Results: Massifquant is an open source solution for Kalman filter isotope trace detection that has been subjected to novel and rigorous methods of performance evaluation. The presented evaluation with accompanying annotations and optimization guide sets a new standard for comparative isotope trace detection. Compared to centWave, matchedFilter, and MZMine2—alternative isotope trace detection engines—Massifquant detected more true isotope traces in a real LC-MS complex sample, especially low-intensity isotope traces. It also offers competitive specificity and equally effective quantitation accuracy.

---

6 Publication authors are Conley CJ, Smith R, Torgrip, RJO, Taylor RM, Tautenhahn, R, and Prince JT
Availability: Massifquant is integrated into XCMS with GPL license > 2:0 and hosted by Bioconductor: http://bioconductor.org. Annotation data is archived at http://hdl.lib.byu.edu/1877/3232. Parameter optimization code and documentation is hosted at https://github.com/topherconley/optimize-it.

Contact: cjconley@ucdavis.edu or jtprince@chem.byu.edu

7.2 Introduction

The most important automated data-analysis step in a typical quantitative -omics XC-MS analysis pipeline is isotope trace (IT) detection. In liquid or gas chromatography mass spectrometry (LC-MS or GC-MS, with either specified as XC-MS) analytes elute with chromatographic separation and are subsequently measured by the mass spectrometer. IT detection is the first and essential step in enumerating the signals of these analytes.

IT detection is a trivial task when performed on data derived from simple mixtures, but can be highly challenging for complex mixtures because there are 1) large numbers of analytes which coelute, many show interlocking or overlapping isotope envelopes; 2) an unknown number of analytes; 3) an abundance of ITs with low signal to noise ratio; 4) significant intensity variation in the signal composing lower abundance ITs due to dynamic range limitations of the spectrometer; and 5) heteroscedastic m/z variance as a function of intensity for most mass spectrometers. Unisotropic m/z variance results in that the data comprising the tails of an IT have larger m/z variance than the data around the mode, and that low abundance ITs have a larger m/z variance than high abundance ITs.

Though difficult to achieve, increasing the sensitivity and accuracy of IT detection software influences the entire downstream analytical pipeline. An example: Vast numbers of peptides
go unidentified in proteomic analyses;\textsuperscript{31} a more sensitive IT detection would allow researchers to track and quantify these peptides, leveraging identifications acquired in other samples. It goes without saying that accurately determining IT boundaries and distinguishing signal from noise improves quantitation results. Furthermore, accuracy in IT detection can also result in more accurate precursor mass estimates and therefore yield an increase in both the number and quality of peptide identifications.

Most IT detection software, such as matchedFilter, relies on the creation of fixed width m/z bins (buckets) to facilitate finding and quantifying eluting analytes. Though bucketing is computationally efficient, for complex data sets it is impossible to find a bin size and position that excludes closely co-eluting ITs while also being broad enough to fully capture the IT of interest. To address this shortcoming, Tautenhahn et al.\textsuperscript{257} developed a software package, centWave, which uses a binless pre-scan to first identify regions of interest composed of centroids. A centroid is a (m/z, intensity) measurement pair at a given time scan of the chromatographic dimension. Once a region is specified, the centroids are then collapsed into a one-dimensional chromatogram and wavelet-based curve fitting is performed to separate closely eluting ITs. The approach is appealing because the initial algorithm identifies zones of interest in a binless way and because the algorithm used for detecting ITs using intensity fluctuation in the time domain is sophisticated. However, in this approach subtle shifts in m/z value are ignored when data are combined into a one dimensional chromatogram. ITs which are very close in m/z or with poor chromatographic profiles may not be properly resolved.

The same year Aberg et al. developed TracMass, a binless IT detection algorithm which fully utilizes m/z information in distinguishing ITs.\textsuperscript{258} TracMass uses a chromatographically
traversing 2-dimensional Kalman filter model (KF)—one dimension focused on m/z values and the other on intensity values—to determine which centroids belong with each extending IT. The decision to incorporate a centroid is made by carefully weighing all previous m/z and intensity evidence of that IT, so misincorporation of centroids is rare as the KFs incorporate more data. Furthermore, the KF accounts for the heteroscedastic variance within the same IT as intensity values change. The KF approach can disentangle even the most closely eluting chromatographic ITs. Furthermore, for the nonexpert user, TracMass requires few user parameters for effective operation.

Despite its apparent promise for IT detection in complex samples, no peer-reviewed publication had compared TracMass performance to leading options until just recently with TracMass2. This is not an isolated deficiency—most IT detection algorithms are not rigorously evaluated because of the difficulty of establishing ground-truth data, especially for lower abundance ITs. Indeed, other compelling binless methods for quantitation may benefit from a similar evaluation as presented here.

Here, we make available an open source implementation of the TracMass algorithm, called Massifquant, and integrate it into the popular XCMS software suit. Like TracMass, Massifquant uses a two-dimensional Kalman filter to quickly, accurately, and adaptively find ITs in highly complex samples without resorting to binning, and its open license (GPL ≥ 2.0) enables further extension and inspection. We indicate how the KF adapts to m/z variance and describe two major modifications which mitigate known limitations of TracMass. We detail novel metrics for evaluating XC-MS IT detection and use these metrics with manually annotated data
to perform a detailed evaluation of Massifquant, centWave, matchedFilter, and MZMine2\textsuperscript{17} performance on different LC-MS platforms.

7.3 Methods

7.3.1 Description of the Massifquant algorithm
Massifquant relies on 2D Kalman filters to identify ITs in XC-MS data. A single KF’s purpose is to track the m/z and intensity coordinates of an IT over the chromatographic dimension. A track is an instance of a KF model, which predicts the existence of a centroid in the next time scan. If the prediction is close enough to a real centroid, it incorporates the real centroid to the track. Closeness is determined by quasi-confidence intervals centered about the prediction. The KF then updates its estimate of the underlying “true” centroid and predicts again. When the signal of the IT disappears (i.e., we have reached the end of a chromatographic IT) the KF will fail to predict a centroid on successive scans and tracking will be terminated.

With many ITs to be discovered, Massifquant manages a host of active KFs. For a given scan, each active KF claims the centroid that best fits its predicted location. Unclaimed centroids trigger new instances of KF tracks in the expectation that these are the beginning of new ITs. The process is then repeated on the next scan until all scans have been examined. In this way, every centroid is either claimed by an existing KF or triggers the creation of a new KF. After an entire sample has been parsed, spurious KFs are discarded based on simple filters for minimum length, intensity, expected m/z deviance, or consecutive missed predictions.

We will describe the Kalman gain to highlight the model’s adaptive nature and how it can be tuned. After the KF predicts a centroid, it refines the prediction by carefully weighting the model prediction error through a modeling device known as the Kalman Gain. This device is
largely a function of (i) the estimation error covariance, which is initialized by the modeler, but evolves over time based on prediction performance; (ii) and the assumed measurement error of the Mass Spectrometer, also defined by the user. So the modeler may tune the Kalman gain based on these parameters. A smaller Kalman gain means that the model prediction, which is based on previous observations, is trusted to be closer to the true centroid location than the newly acquired observation. The default settings of Massifquant create a Kalman gain that places more trust in early acquired observations (i.e. the first 4-30 scans) as illustrated in Figure S1 in the supplementary materials. The idea is to find the IT’s location quickly and not deviate once it has been found; the default works for a variety of situations, but can also be tuned to a particular dataset. The fact that the KF continuously adapts its centroid prediction estimates based on the information it has previously amassed and the variance it encounters makes it an effective tool for identifying ITs with their own specific heteroscedastic variance. For a more mathematical discussion, an introduction to the theory behind the discrete Kalman Filter/Gain are described in Welch and Bishop and section 2 of the supplementary materials.

Massifquant implements most of the core of the TracMass algorithm; however, it is difficult to determine how much the two algorithms differ since the latter’s source code is not available. There are a few known major differences. The initialization of the $P$ is likely different. Moreover, the intensity component of the Scheffe-type quasi-confidence intervals—used to classify whether a next-scan centroid belongs to a KF prediction—was not found to be sufficiently discriminatory. Massifquant only uses the m/z dimension to determine a successful prediction. Retaining the intensity estimation in the KF does seem to aid in resolving competing KFs that claim the same centroid (by virtue of comparing their two dimensional prediction distances).
Massifquant also implements a function to ensure continuity of identified ITs that is not found in TracMass (discussed in section 3 of the supplementary information). We found that a KF will periodically lose the position of the IT, stop tracking it en route, triggering a new KF track which will finish estimating the IT’s other data points (see supp. file Figure S2). Since each KF track corresponds to an IT, we call the undesirable phenomenon “segmentation”. The segmentation problem was addressed by an ad-hoc t-test comparing the m/z locations between these problematic KF. The conservative test combines many of the segmented tracks into a unified IT.

A more thorough description of the Massifquant implementation is given in the supplementary material (see the section ”Reimplementing the Kalman filter model”). The supplement highlights some differences with TracMass and a discussion of the logic behind specific design decisions. The description will be useful to anyone seeking to modify or extend the algorithm. Massifquant was written in C++ and has been integrated into the XCMS pipeline available through Bioconductor. It plays the same role as centWave, matchedFilter, or MZMine2’s isotope trace detection algorithm in the differential analysis workflow.

7.3.2 Annotation

7.3.2.1 Data sets
We chose two very different LC-MS data sets to assess IT detection flexibility. The first annotated data set, MM14, is a subset from a UPLC-ESI-QTOF analysis of 14 plant metabolites resulting in 46 annotated ITs. The centWave developers originally showcased their method of parameter optimization on the entire data set, and its provenance is detailed in Tautenhahn et al.
The second data set, MOUSE, is one fraction from a larger mouse brain phosphoproteomic analysis. Briefly, 408.8 mg of brain tissue was homogenized/boiled in SDS-lysis buffer and clarified. Proteins were then digested and peptides purified following the FASP protocol to yield an estimated 7.3 mg of peptides. 25 mg of Titanspere TiO2 beads (GL Sciences) were used to enrich for phosphorylated peptides. 3M Empore Anion Exchange disks were packed into a 200 l pipette and Britton & Robinson buffer was used to elute at pH 11 (the fraction termed ’MOUSE’ in this work), 6, 5, 4, and 2. MS analysis was performed with an LTQ-Orbitrap XL fed by an Eksigent NanoLC UHPLC system. A Nano Acquity (1.7m, 130 C18 bead BEH, 75m m x 150mm) column run at 375 nL/min in a linear gradient from 2.5% to 10% ACN (with water and 0.1% formic acid as the second buffer) for 60 minutes, then to 28% ACN for an additional 220 minutes. The complete raw file is available upon request, and virtually all parameters may be accessed using the cross-platform unfinnigan software (see https://code.google.com/p/unfinnigan/). The relevant parameters are: MS1 data collected between 375–1800 m/z at 60,000 resolution with an MS/MS data dependent scan collected after each MS scan. The section chosen for hand-annotation generally spans retention time 5429.5–7306.2 seconds and 600.0003–637.3923 m/z. In total, this area contained 589 annotated ITs which show variation in length, shape, and variance.

7.3.2.2 Data annotation

The MOUSE and MM14 datasets were manually annotated to be used as ground truth for assessing the automated IT detection abilities. A tuned LC-nanoESI system is capable of producing consistent chromatographic IT shapes. However, when running complex samples, even on the best tuned system, fundamental dynamic range limitations will unavoidably produce IT shapes that are far from ideal. The lack of characteristic IT shapes among lower abundance
ITs, the number of overlapping ITs (in m/z and time), and their sheer number and density makes manual annotation difficult. For the MOUSE data, any IT that did not exceed a maximum intensity of $1 \times 10^5$ was ignored to preserve the integrity of the annotation.

Because IT annotation in complex data sets is challenging, we established guidelines for what is called a true IT. These guidelines consider within-IT and between-IT characteristics to ensure the best annotation possible. To be defined as an IT, a series of centroids should typically exhibit the following properties:

Within 1. The m/z error variance structure is influenced by intensity. Toward the tails of an IT, the m/z observations show mostly symmetric and increasing deviations from the mean. The body and apex centroids deviate less. From a bird’s eye view (i.e., looking down the intensity axis), the m/z-time projection has the shape of a string fraying at the edges.

2. The collective centroids should have a chromatographic IT shape. Dramatic oscillations in intensity from scan to scan could disqualify an annotation.

Between 1. The detected ITs should have approximately the same m/z ppm variance.

2. Within an isotopic envelope, ITs should have very similar mode and shape, although length typically varies.

In each case, great effort was made to balance the benefits of the systematic application of these rules with human judgment. Each IT was individually annotated (based on all criteria) and then wrapped into appropriate isotopic distributions where possible.

We executed this annotation scheme on the MM14 and MOUSE data sets using Topp-View as follows: From a global 2-D view, the annotator identified mass traces satisfying mentioned
properties. After zooming, a 3-D inspection confirmed similar chromatographic length and shape for a given isotopic distribution. Once confirmed, the IT’s centroids were selected and collectively saved into an .mzML file. Candidate mass traces that did not sufficiently satisfy all the criteria, but still had some resemblance to an IT, were labeled as questionable and saved as .mzML files; these were excluded from the algorithm performance analysis since they were deemed liable to interfere with true algorithmic specificity and sensitivity. Objectively determining an IT’s chromatographic boundaries is difficult, especially since there is so much diversity among IT shape and length. Generally, we tried to include as much of each IT tail as possible and to be as consistent as possible across each data set.

7.3.3 Performance Evaluation
Different algorithms select different portions of an IT when attempting to identify ITs (any attempted IT classification we call a candidate). Because the extent and location of the mapping from a candidate to the true IT may vary widely, gauging the success of a candidate can be challenging. For example, a method that identifies 30 centroids directly in the middle of the high intensity region of an IT should be given more credit than one that identifies 35 centroids but that are all in the very low intensity tail region. In another example, credit should be given to an algorithm that successfully captures an entire IT with three distinct candidates, but it should not receive as much credit as an algorithm that identified the IT with a single candidate. These examples motivated the development of two ways of examining success: at the IT-level and at the entire sample-level.

7.3.3.1 Isotope trace-level evaluation
Classifying the success of an algorithm at the IT-level requires the classification to be general enough to handle a variety of IT shapes and yet still be precise. To classify the successful
identification of an IT, we defined metrics that consider how a candidate’s centroids individually contribute to the overall intensity of the annotated IT, namely, the true area under the curve ($AUC_A$). The centroids clustered into a candidate are either true positives, false positives, or false negatives. Restricting attention to the true positives, a candidate’s true area under the curve is denoted as $AUC_{TP}$. Naturally, a candidate’s relative correct identification of an IT within the context of intensity is defined to be $\alpha = \frac{AUC_{TP}}{AUC_A}$. Now, an algorithm is said to sufficiently identify the $i_{th}$ annotated IT if $\alpha_i \geq 1 - r$, where $0 \leq r \leq 1$. For the following analysis, we took $r = 0.5$ because requiring a candidate to capture more than 50% of an IT’s total intensity ensures that the main body of a IT has been identified, while still allowing for differences in opinion on exact IT boundaries. In short, this criterion abstracts away the difficulty of varying shapes and algorithmic-selection bias.

Conversely, the false positive and false negative centroids contain precise information as to where a candidate is accurate and by how much. To be clear, the AUC quantitation error is taking evaluation precision beyond classification. Let AUC be the quantification reported by the algorithm, which includes true and false positive centroids alike and excludes false negative centroids. Then the AUC percent error is simply $\epsilon = \left| \frac{AUC_{A} - AUC_b}{AUC_{A}} \right| \times 100\%$. Dramatic variation in IT intensity motivated the percent error representation.

Another issue is that true negative ITs are impossible to define. So an algorithm’s IT-identification accuracy was measured by the commonly used metrics of precision and recall (sensitivity) for information retrieval. Isotope trace sensitivity ($s_f$) is the number of ITs correctly identified by the algorithm divided by the number of true ITs. Isotope trace precision ($p_f$) is the
number of ITs correctly identified by the algorithm divided by the number of algorithm-claimed ITs. High sensitivity means the algorithm successfully identifies most true ITs, while high precision is a measure of identification reliability. The harmonic mean of these is the $F_1$ score, $2 \frac{s_p p_f}{s_f + p_f}$; it summarizes the overall identification performance.

### 7.3.3.2 Sample-level evaluation

Finally, sample-level metrics allow us to define how much of the entire sample AUC was correctly identified without regard for individual ITs. It is a way to quantify the level of intensity information found by an IT detection without regard to how the centroids are actually clustered into ITs. The sample sensitivity is defined as $\frac{\sum_i AUC_{T_i} p_i}{\sum_j AUC_{A_j}}$. This is the total algorithm-identified true raw intensity divided by total true raw intensity. On this global level, a true negative can be defined as the sample noise, or the centroids that don’t contribute to any real ITs. Thus, the sample specificity equals $\frac{\sum_i AUC_{T_i} N_i}{\sum_j AUC_{F_j} p_j + \sum_k AUC_{N_k}}$. This taken to be the total correctly algorithm-ignored raw intensity (true negative signal) divided by total noise raw intensity of the sample (including false positives of the algorithm). These last two metrics are useful as a global measure of accuracy in contrast to the IT-specific accuracy in the preceding metrics.

### 7.3.3.3 Evaluation by IT type

An evaluation should indicate how certain IT types influence performance. Simpson’s paradox further motivates an evaluation by type since conclusions based on the aggregate annotation are sometimes reversed when analyzed by type. We classified ITs by intensity, ppm error, and length. Annotated ITs were grouped by the variable of interest into 8 percentile categories {[0, 12.5%), [12.5%, 25%) . . . , [87.5%, 100%)}. For example, the longest IT was categorized in
The recall was computed for each category; precision was approximate because mapping the algorithm-identified ITs to the right annotation-based category was not always right. For instance an algorithm-identified IT length might be shorter or longer than the annotation length and the mapping can only be corrected if the IT identification is correct.

7.3.3.4 Optimization

With the goal of maximizing the F1-score, we optimized parameters for the two algorithms on each dataset. Initial values for centWave on MM14 were selected from the paper Tautenhahn et al.\textsuperscript{257}; the manual annotations provided a baseline of minimum IT length, height, and ppm deviation. Where prior knowledge was absent, liberal parameter grids were explored for parameters like \textit{snthresh} for centWave, or \textit{criticalValue} for Massifquant. Paired parameters, or parameters that were thought to have interactions, were explored simultaneously in two dimensions. For instance, the (min, max) IT length form a natural pair and exhibited interactions in F-score performance for centWave. The most important parameters for both algorithms, \textit{(snthresh, ppm)} in centWave, and \textit{(criticalValue, consecMissedLim)} in Massifquant were searched simultaneously. Their respective F-score surface plots exhibited near-concavity, a desirable property for parameter tuning. It appears unique to Massifquant that all F-score surface plots had near-concavity. The optimizations were conducted with R (http://www.r-project.org) and Matlab scripts (MATLAB version 7.14.0.739, The Mathworks Inc., Natick, Massachusetts) scripts and detailed procedures to reproduce results are available on GitHub (see Availability).

Other details of the optimization are included in the supplementary file. Table 1 compares centWave performance on MM14 based on reported optimized parameters from the original centWave publication and the optimized parameters resulting from this new evaluation. The two different evaluation settings yield similar parameters and \textit{F}_1-scores, suggesting this new
annotation and evaluation effort is valid. For matchedFilter and MZMine2, all combinations of
the suggested ranges for each parameter were exhaustively evaluated (see supplement for the full
list).

**Table 7.1 centWave optimization on MM14 improved with identification performance and the parameters are in the
same vicinity.**

<table>
<thead>
<tr>
<th>version</th>
<th>ppm</th>
<th>snthresh</th>
<th>peakwidth</th>
<th>peakfilter</th>
<th>F₁-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>30</td>
<td>2</td>
<td>(5,10)</td>
<td>(2,400)</td>
<td>0.8936</td>
</tr>
<tr>
<td>our evaluation</td>
<td>18.4</td>
<td>2.5</td>
<td>(3,11)</td>
<td>(1,511)</td>
<td>0.9438</td>
</tr>
</tbody>
</table>

7.4 Results

7.4.1 Overall Evaluation

As detailed in the methods section, we developed an independent, open-source implementation
of Aberg et al.’s TracMass algorithm, and call it ‘Massifquant’. The algorithm uses 2-
dimensional Kalman filters to adaptively find chromatographic ITs in the m/z domain without
bucketing the data. We compared Massifquant’s ability to sensitively and accurately find ITs
with centWave, a sophisticated and well-known algorithm used in the XCMS platform for label-
free IT detection, matchedFilter, the original binning-based XCMS method for IT detection, and
MZMine2, a non-XCMS platform for MS data processing.
We manually annotated ITs in two data sets, chosen to have different characteristics, following a set of rational guidelines. The MM14 data set is a run of 14 plant metabolites on a lower resolution UPLC-ESI-QTOF. The MM14 reveals the performance of an IT finder under close to ideal circumstances (viz. low sample complexity, good signal-to-noise, good chromatography).
The MOUSE sample was run on an Orbitrap mass spectrometer and is typical of many highly complex proteomic analyses. While chromatographic IT shapes are smooth for high abundance ITs, the intrinsic dynamic range limitations result in greater m/z and intensity variability for lower abundance analytes. The heterogeneity of IT sizes and shapes encountered in the MOUSE data is ideal for discovering the limitations of an IT detection algorithm.

Figure 7.1 shows that Massifquant reported uniformly higher sensitivity values than centWave and the t-test union of segmented ITs improves Massifquant performance on MOUSE. As for

Figure 7.2: A comparison of log-transformed percent quantitation errors (log ϵ) for successfully identified ITs. Massifquant outperforms centWave’s quantitation error on both data sets.
identification reliability, precision was in the same neighborhood for both datasets, yet centWave shows higher sample specificity in MOUSE since it rarely found a false IT. Massifquant exhibited a better F1-score on MOUSE since it identified substantially more ITs than centWave.

![Graphical representation of manually annotated ITs on MOUSE data set and detected ITs, for centwave and Massifquant. Correctly identified ITs are color-coded according to the percent quantitation error (ε): dark blue < 10%, aqua < 20%, green < 40%, orange > 40%. False ITs are labeled in red; all other noise was excluded. ITs missed by the algorithm (i.e., false negatives) are labeled black.](image)

Figure 7.3: A comprehensive view of manually annotated ITs on the MOUSE data set and detected ITs, for A) centwave and B) Massifquant. Correctly identified ITs are color-coded according to the percent quantitation error (ε): dark blue < 10%, aqua < 20%, green < 40%, orange > 40%. False ITs are labeled in red; all other noise was excluded. ITs missed by the algorithm (i.e., false negatives) are labeled black.
Both algorithm’s MM14 performance is effectively equal for all metrics but sensitivity. The matchedFilter algorithm was only able to identify 33 of the 589 ITs in the MOUSE dataset after optimization over 215 parameter settings. MZMine2’s best performance was worse, with only 20 isotope traces correctly identified under optimal parameter settings (see supplement). Because matchedFilter and MZMine2 perform so poorly compared to centWave and Massifquant, we omit the results from the charts in this paper.

Comparing algorithms’ quantitation accuracy is controversial because defining IT boundaries is not clear-cut and in this analysis most error comes from the tails—knowledge afforded because of the evaluation criterion. No statistical test comparing the two algorithm’s was done since the spatial components, length, shape, m/z variance, etc. likely create dependence among ITs. Nonetheless, Figure 7.2 illustrates that Massifquant and centWave quantitation errors are generally in the same small neighborhood.

7.4.2 Evaluation by IT Type
An evaluation is incomplete without identifying what types of ITs were missed within certain types of samples. For example, both algorithms are perhaps equally excellent at detecting ITs in a simple sample like MM14 with high signal-to-noise (see supp. data Figure S3). On the other hand, Figure 7.3 shows that Massifquant excels at finding low-intensity type ITs in the MOUSE complex sample and quantifies them very well, whilst these are not identified by centWave.

The “Evaluation by IT Type’ strategy’, described in section 2.3, addresses whether the high number of low-intensity ITs relative to high-intensity ITs in the MOUSE data unfairly benefited Massifquant in aggregate statistics (viz. F1-score). Figure 7.4 summarizes the results of IT-typed performance for characteristics thought to vary widely within MOUSE. centWave’s IT
sensitivity improves as the intensity increases and the estimated ppm error decreases, both in a linear fashion. *Massifquant’s* sensitivity varies little across all categories, irrespective of the variable, and without a doubt outperforms centWave. With respect to IT precision, the effect of each variable seems present for both algorithms. Both have similar approximate precision results. Not surprisingly, Massifquant shows improved precision as length, narrowness, and max-intensity increase.

**Figure 7.4**: Isotope trace detection performance by quantiles for IT characteristics of the MOUSE data set. The left-most percentile bins generally represent the hardest cases for IT detection algorithms (short, low intensity, broad ITs) while bins on the right are generally easier (long, high intensity, narrow ITs). The sensitivity panel is at the IT-level.
7.5 Discussion & Conclusions

In Massifquant, we have implemented an open-source Kalman filter-based IT detection algorithm based on Aberg et al.\textsuperscript{258} We have evaluated its performance using two manually-annotated data sets, and compared the performance of Massifquant with centWave, a wavelet-based IT finder, and matchedFilter and MZMine2, binning-based IT finders. A protocol for how IT detection algorithms should be evaluated has not yet been established, so we first discuss the evaluation process; then, we address algorithmic performance and suitability for use, and finally conclude with some thoughts about the use of m/z information in MS IT detection generally.

7.5.1 The evaluation process

Comparative evaluation of algorithms in MS-omics is often lacking Smith et al., and Zhou et al. recently suggested that the quantitative evaluation of IT detection algorithms is long overdue.\textsuperscript{32,260} We believe the general lack of evaluation is related to the difficulties associated with creating data sets to effectively test these algorithms and also to a lack of clear and explicit metrics for assessing success. In order to facilitate further efforts in this area, we discuss some of the challenges and successes we met using a manually annotated data set approach.

Hand-annotation, especially of low abundance ITs, is extremely challenging. It requires concerted effort over a long period of time. The authors spent several weeks of dedicated effort in order to annotate the two data sets, and the MOUSE data set is only a small subset of the complex LC-MS sample from which it was derived. Despite our best efforts to be accurate and consistent, we conclude that the manual annotation process is still somewhat subjective. Indeed, we simply had to exclude the evaluation of ITs below a certain threshold because we felt human judgment was inadequate for the task. Despite these challenges, the annotation data itself is a
useful model for future validation efforts. Moreover, it contains isotopic-level information that could be of use in other projects.

We validated the manual annotation efforts through a holistic visual inspection (see Figure 7.3 for example) and analysis of histograms of ppm deviation (see supp. Figure S4 for example) to ensure that there were no outliers. So, despite the inherent difficulty of manual annotation, we conclude that the endeavor was largely successful. Several aspects of the process are worth considering in more depth: 1) we used semi-rigid guidelines for annotation that we believe worked well across a variety of ITs with different characteristics. We could have generated and applied very strict rules for annotation at the outset, but this may have resulted in even worse systematic bias considering the highly variable ITs we encountered. The proposed guidelines should serve useful for future annotation efforts. 2) We used a single annotator for both data sets to eliminate person-to-person variability in the interpretation and application of IT criteria. However, tools for community sourcing annotations would be an interesting solution and has been already been discussed in genomic contexts.\textsuperscript{269} 3) We used ToppView, the MS viewer associated with OpenMS, to help us find and annotate ITs.\textsuperscript{25,267} Additional add-ons such as color-coding and flagging of already-annotated ITs and producing a community based validation would also improve the annotation process.

Among the previous efforts to evaluate IT detection algorithms, we found that most of them focused solely on questions of identification, but lacked in detail of what constituted an ’identified’ IT. For IT detection, the identification criterion is critical for fair evaluation—and we additionally argue that the evaluation should probe quantitation accuracy if possible. We evaluated identification at IT and sample levels, and also calculated the percent quantitation error
for each IT. The precisely defined metrics may now be more easily employed, modified, or improved.

This multi-metric evaluation exposes two risks other evaluations take when relying purely on the F1-score. 1) Precision values show that Massifquant does at least as well if not better at IT identification reliability for MOUSE at low intensity. However, the sample specificity, along with Figure 7.3, provide stronger evidence that centWave effectively discriminates low-intensity non-ITs better than Massifquant. Hence, precision and consequently the F1-score can be misleading. To our knowledge, this is the first evaluation that has proposed a true specificity measure for IT detection, which helps avert wrong conclusions. 2) By our evaluation standards, and likely others, accurate quantitation does not always imply a favorable IT detection F-score and vice-versa. On the MOUSE dataset centWave ignores many low-intensity ITs, giving it a low F-score; however, the ITs that it does identify are generally quantitatively accurate with a median $\epsilon = 8.663\%$. Thus, quantitative accuracy is somewhat distinct from IT detection sensitivity or precision.

### 7.5.2 Algorithm performance

On the simple data set MM14, Massifquant showed similar performance to centWave. On a highly complex sample, MOUSE, Massifquant performed much better. In particular, Massifquant excels at finding ITs with a variety of characteristics such as differing intensity, widths, and lengths. Massifquant outperforms centWave in IT detection sensitivity across every size and shape of ITs in the complex sample tested. As for reliability, Massifquant is competitive with centWave with the exception that it finds more false low-intensity ITs; the excess false positives and multi-modal artifacts are two deficiencies of Massifquant which can
complicate downstream analysis in sample-to-sample comparisons. Future extensions of Kalman Filter IT detection will need to make intensity estimation more robust. An attempt to combine centWave’s wavelet intensity estimation with Massifquant has not proven to be effective (see supplement section 4). In spite of these deficiencies, both algorithms reported similar quantitation accuracy for the quantified ITs; Massifquant just found far more ITs.

A possible objection to our general comparison is that a large number of small ITs might bias the evaluation in Massifquant’s favor. However, Figure 7.4 removes any suspicion of unfair advantage; even if low-intensity or very broad ITs (e.g. first four bins) were removed from the analysis, Massifquant still identifies ITs better on the MOUSE data set.

As shown in Figure 7.1, our effort to address the problem of IT segmentation with Massifquant was successful—on the MOUSE and MM14 data set, the precision increased from 0.7391 to 0.7894 and 0.9185 to 0.9355, respectively. However, some ITs were erroneously combined (see supp. data Figure S2). For algorithmic simplicity, future efforts should attempt to address the IT segmentation problem from within the framework of the Kalman filter. Ideally, such an approach would also be more effective than the ad-hoc method we applied in this study to treat IT segmentation.

7.5.3 Ease of use
Massifquant parameters can be readily optimized through visual confirmation instead of score-based methods (e.g. f-score) that require an annotation. Visual optimization is more time efficient, intuitively simple, and almost as accurate. Similar in purpose to Tengstrand et al.,261 the visualization tools at https://github.com/topherconley/optimize-it illustrate precise changes in IT detection induced by differing parameter input. The documentation offers a step-by-step guide
how to optimize Massifquant to new data sets, especially controlling the number of false positives. Further, the score-based method shows a concave f-score surface when varying Massifquant’s parameters, indicating a very predictable parameter behavior (see supp. data Figure S5, S12, S13, and S14). Massifquant’s appeal is due, at least in part, to the fact that several internal KF parameters are learned from the data—in an initial prescan, and then later for each individual IT being tracked.

Massifquant operates on centroided MS data, which means it can analyze data taken in centroid mode or profile mode (after centroiding), whereas algorithms requiring profile data cannot

---

**Figure 7.5:** Real-world application test. Massifquant identifies differentially expressed ITs between wild type (WT) vs. knock-out (KO) conditions in the faahKO dataset for (A) trivial cases and (B) non-trivial cases.
operate on centroid data because the centroiding process is not readily reversible. Further, running Massifquant is as easy to run and modular as other XCMS IT detection options. The same differential abundance (DA) workflow applies. Figure 7.5 illustrates a Massifquant-based DA analysis in on the FAAH knock out LC/MS data set,\textsuperscript{270} (see http://bioconductor.org/packages/devel/data/experiment/manuals/faahKO/man/faahKO.pdf for details).

7.5.4 The use of m/z information in IT detection
Can the success of Massifquant on a complex sample be generalized? ITs in a highly complex sample—particularly low abundance ITs—are different from ITs derived from a simple mixture: limitations in a mass spectrometer’s dynamic range produce much greater intensity variability for ITs from a complex sample. Because of this, at least for mid-to-high mass accuracy/resolution mass spectrometers, m/z measurements will tend to be far more helpful at distinguishing closely eluting species than IT shape. Indeed, we found that Massifquant performs at a high level because of its m/z estimation (despite extremely poor intensity estimation). Most IT detection algorithms focus on IT shape, but we suggest that on highly complex samples an algorithm should be focused mainly on subtle changes in m/z. Algorithms that bin data from closely related ITs in order to do IT shape analysis lose the richest information available for distinguishing those ITs. Distinguishing convolved isobaric compounds and near isobaric compounds will, of course, require chromatographic IT shape analysis, but new algorithms will likely see the greatest improvement gains by working to fully utilize the m/z information found in closely eluting analytes.
Chapter 8  Metriculator: quality assessment for mass spectrometry-based proteomics

Author’s Note: This chapter describes novel software for tracking, monitoring, and automating performance metric generation for mass spectrometers. While the discussed application is proteomics, the software is applicable to all MS based omic studies. I wrote the archival, data extraction, and data management algorithms, created the custom visualization, and built much of the web interface. These results were published in Bioinformatics.²⁷¹⁷

8.1 Abstract

Summary: Quality control in mass spectrometry-based proteomics remains subjective, labor-intensive and inconsistent between laboratories. We introduce Metriculator, a software designed to facilitate long-term storage of extensive performance metrics as introduced by NIST in 2010. Metriculator features a web interface that generates interactive comparison plots for contextual understanding of metric values and an automated metric generation toolkit. The comparison plots are designed for at-a-glance determination of outliers and trends in the datasets, together with relevant statistical comparisons. Easy-to-use quantitative comparisons and a framework for integration plugins will encourage a culture of quality assurance within the proteomics community.


8.2 Introduction

As omics-level experiments increase in size and complexity, assessing the quality of a dataset can be a laborious undertaking. This is particularly true of mass spectrometry (MS)-based...

² Publication authors are Taylor RM, Dance J, Taylor RJ, and Prince JT
proteomics, where the spectrometer and associated chromatography exhibit variable—and
sometimes erratic—performance. Researchers would like to analyze more samples and in
greater depth (i.e. fractionation), but maintaining high quality across the set—and knowing that
the set is of high quality—is a mounting challenge in proteomics.

Typically, highly trained technicians spend significant time adjusting capillary plumbing,
working to achieve stable nanoelectrospray and calibrating and tuning the mass spectrometer.
The quality of a large analysis is then assessed by visual inspection of the 2D or 3D ion trace.
Assessing reproducibility is particularly challenging when datasets involve runs collected over a
period of weeks or months. Furthermore, human assessment of quality is both time-consuming
and a potential source of bias. Software tools to aid in quality assessment are needed and can
improve confidence in published proteomic datasets.

An extensive set of quality/performance metrics was introduced by NIST to begin to assess data
quality in MS proteomics. These 284 metrics include measures of chromatographic
performance, ion source stability, ion signal intensity and data-dependent sampling efficiency.
These measures can significantly augment manual interpretation of data quality, but their utility
depends on contextual comparisons between datasets. Comparing metrics over time is the key to
leveraging them for quality assessment. Subsequent work has attempted to expand metrics
across vendor platforms, and to provide curated metrics, such as those demonstrated by
QuaMeter, indicating the value of metrics to the proteomics community.

A similar suite of software designed to monitor lock mass and quality control evaluation at the
instrument and identification level, called SIMPATIQCO, was recently released. It provides a
similar web-based interface designed to assist instrument operators in monitoring quality control
samples, yet it lacks interactive graphing and comparison capabilities, which enable applicability
to any questions of performance differences, as well as augmentation to existent workflows.
Additionally, there is a commercial product, MassQC,\textsuperscript{276} which provides a utility for submitting
metric information for longitudinal comparison, but does not provide automation. There remains
a need for an open-source configurable method for tracking and comparing performance metrics,
as well as integration into a proteomics workflow.

8.3 Software
Here we introduce Metriculator, an easily installable database backed web service that generates,
stores and compares metrics across datasets for quality control purposes. We also provide
archival features to facilitate automatic metric generation of the NIST metrics and workflow
integration. This package is meant to serve as a framework for an automated workflow
customizable by each research group. We chose Ruby to enable users to easily extend the
framework through our integration plug-in setup; Ruby is easy to learn,\textsuperscript{177} boasts a large number
of off-the-shelf utilities for web programming, and is gaining traction in the scientific
community.\textsuperscript{277}

8.3.1 Implementation
Metriculator is cross-platform, tested on *NIX systems and Windows, and only requires an
installation of Ruby (version ≥1.9) to function. The interface is built on Rails, the popular web
framework, and provides interactive graphs through the HighCharts library.\textsuperscript{278}

8.3.2 Metric generation
Metriculator automates generation of the 284 NIST performance metrics from LC-MS run
*RAW files and stores them in a database to ensure that relevant meta-information can be
compared over time even if raw data are lost. The NIST metrics are generated by the NIST MSQC software, which is platform specific for Windows, as detailed by NIST.\textsuperscript{273}

8.3.3 Web interface
The browser interface is designed for easy access to the stored metric information. Through it, a user can access any of the meta-information about a run and can generate a comparison of all metric values between two lists of msruns. These comparisons provide graphs designed to provide at-a-glance evaluation of the metric information. The website also provides for email based alerts, customizable via a QC\_alerts configuration file, which specifies a notification threshold, in deviations from the historic mean for each metric, to trigger an alert email, notifying technicians of instrument problems immediately. This utility could easily be expanded to hook into any notification systems.

8.3.4 Visualization
The graphs incorporate a time-rendered plot of the data points for each set of msruns, as well as a visual comparison of the two populations by beanplot.\textsuperscript{279} Beanplots provide a compact visual summary and comparison between two distinct populations without sacrificing visualization of potentially interesting individual datapoints. They consist of two vertically plotted density plots, with a univariate plot of individual datapoints. Each plot provides at-test comparison, as well as an immediate visual summary of any trends or significant differences in the comparison sets of msruns. These plots are generated dynamically via Javascript and are interactive to enable simple identifications of anomalous metric values (Figure 8.1).
Ms-archiver—integration with a workflow

Metriculator is ideally integrated into a workflow to ensure that metrics are generated for every run. Metriculator allows for automated data off-loading/backup and metric generation on run completion, to reduce loading of data acquisition systems. Communication between computer nodes is accomplished by a simple file system-based queue; the automation framework only requires access to a shared storage location, common to most laboratories. Complete automation can be achieved through the use of a cascading set of settings files in the archival directory and use of the integration capabilities provided.

In its current scope, Metriculator fills a niche role not covered by existing Laboratory Information Management Systems and analysis software (e.g. CPAS). Although other software provides some utility, the ease of use and plotting capabilities of Metriculator provide more intuitive investigation of metric datapoints, as well as an extensible framework for pipeline management. An open customizable code-base allows others to expand on the software to suit their needs through integrating their own automation tools to the integration plugin framework provided by Metriculator, thereby encouraging metric adoption in the proteomics community.

\textbf{Figure 8.1: Visualization Plot example.} It is representative of those generated by the software, showing the combination of a bean plot with the associated time plot for a Visualization Plot, representative of those generated by the software, showing the combination of a bean plot with the associated time plot for a single metric, comparing 7 values with 13 values in a second dataset.

8.4 Ms-archiver—integration with a workflow

Metriculator is ideally integrated into a workflow to ensure that metrics are generated for every run. Metriculator allows for automated data off-loading/backup and metric generation on run completion, to reduce loading of data acquisition systems. Communication between computer nodes is accomplished by a simple file system-based queue; the automation framework only requires access to a shared storage location, common to most laboratories. Complete automation can be achieved through the use of a cascading set of settings files in the archival directory and use of the integration capabilities provided.

In its current scope, Metriculator fills a niche role not covered by existing Laboratory Information Management Systems and analysis software (e.g. CPAS). Although other software provides some utility, the ease of use and plotting capabilities of Metriculator provide more intuitive investigation of metric datapoints, as well as an extensible framework for pipeline management. An open customizable code-base allows others to expand on the software to suit their needs through integrating their own automation tools to the integration plugin framework provided by Metriculator, thereby encouraging metric adoption in the proteomics community.
Chapter 9  Mspire-Simulator: LC-MS Shotgun Proteomic Simulator for Creating Realistic Gold Standard Data

Author’s Note: This chapter represents the first software capable of simulated mass spectrometry data useful for establishing comparisons between data analysis algorithms. I contributed the algorithms for charge state estimation. These results were published in the Journal of Proteome Research.

9.1 Abstract

The most important step in any quantitative proteomic pipeline is feature detection (aka peak picking). However, generating quality hand-annotated data sets to validate the algorithms, especially for lower abundance peaks, is nearly impossible. An alternative for creating gold standard data is to simulate it with features closely mimicking real data. We present Mspire-Simulator, a free, open-source shotgun proteomic simulator that goes beyond previous simulation attempts by generating LC-MS features with realistic m/z and intensity variance along with other noise components. It also includes machine-learned models for retention time and peak intensity prediction and a genetic algorithm to custom fit model parameters for experimental data sets. We show that these methods are applicable to data from three different mass spectrometers, including two fundamentally different types, and show visually and analytically that simulated peaks are nearly indistinguishable from actual data. Researchers can use simulated data to rigorously test quantitation software, and proteomic researchers may benefit from overlaying simulated data on actual data sets.

---

8 Publication authors are Noyce AB, Smith R, Dalgleish, J, Taylor RM, Erb KC, Okuda, N and Prince JT
9.2 Introduction

A single liquid chromatography−mass spectrometry (LC-MS) run is inherently capable of quantifying upward of 100,000 peptides.\(^{31}\) Unfortunately, in a typical analysis, the majority of this data is discarded due to difficulties in identifying and accurately picking chromatographic peaks, especially those of lower abundance. Increasing the accuracy of peak picking results in the detection of more features that can be compared across runs. More accurate peak picking can also influence mass estimates and therefore yield an increase in the number and quality of identifications.\(^{29}\) It ultimately simplifies cross-run comparisons of feature abundances and increases the overall accuracy of those quantitative comparisons. In other words, peak picking quality influences the entire downstream analysis.

For these reasons, it is undoubted that the most important step of a proteomic workflow is feature detection, for which many algorithms exist.\(^{29,255,281}\) However, very little has been done to fully test or compare the performance of these algorithms. In large part, this is due to the challenging nature of creating gold standard data. Fully annotating actual complex proteomic data sets, or even small portions, is extremely time-consuming, difficult, error prone, and subjective. Because MS/MS annotation is rare for small peaks and because they have intensities near the signal-to-noise threshold, accurate human annotation of small peaks in a complex sample is very likely impossible.

Simulation is routinely used in related fields when gold standard data is difficult to come by (e.g., systems biology network simulation)\(^{282}\) or the cost of performing each experiment is high (simulated ion movement in MS fields).\(^{283}\) For quantitative mass spectrometry, an attractive alternative to using hand-labeled data sets is to simulate actual data using noise parameters
derived from experimental data. An ideal simulator would generate data sets where all aspects of the data are known, the various noise components are adjustable, and the peak characteristics conform to those found in biologically derived data sets. Such data sets would be invaluable for comparing algorithms because accuracy can be comprehensively and quickly ascertained programmatically. The speed of this feedback will also aid in the creation of new, more sophisticated algorithms.

Because simulators can produce fully defined peaks of any size, data sets produced by simulation are particularly well suited to test algorithms for their ability to detect and accurately quantify small LC peaks. Small peaks are highly desirable targets for identification and quantitation because: (1) seminal biological events may occur at low quantities (e.g., upstream signal transduction), (2) a change in state to low quantity may be as significant as an increase in quantity, (3) post-translational modifications may manifest themselves as a drop in the unmodified peptide’s concentration, and (4) lower abundance peaks constitute the majority of peaks in an LC-MS run and these are inaccessible by current MS/MS regimes. By quantifying low abundance peaks, proteomic and individual protein coverage may improve, and intraprotein variation can be tracked.

Many proteomic workflows allow users to examine their experimentally derived fragmentation spectra alongside a representation of the theoretically matched spectrum (i.e., an MS/MS fragmentation view). With a simulated LC/MS data set in hand, a somewhat analogous view could be generated for the user where simulated MS1 data is layered on top of actual data. This view would encourage a researcher to examine their MS1 output to reconcile what they can observe with what they expect to observe. A peak that went unidentified may still be present,
and researchers would then know where to look within their MS1 data. Alternatively, a peak that
should have been present may be absent prompting researchers to simulate data with conjectured
post-translational modifications in an effort to locate the modified peak. Simulated data sets
have the potential to augment the traditional proteomics workflow in which researchers often
neglect to thoroughly examine their MS1 data.

While previous MS1 proteomic simulators\textsuperscript{284,285} have been created, a simulator that mimics the
intensity and m/z variance found in real data sets is critical for testing peak-picking/ quantitation
algorithms. Here we present a full featured LC/ MS shotgun proteomics simulator, Mspire-
Simulator, which generates peptide peaks with realistic m/z and intensity variance and elution
profiles. Machine learning is used to generate peaks with a realistic retention time distribution as
well as peak heights reflecting peptide ionization efficiency.

9.3 Methods
Mspire-Simulator takes as its input FASTA files containing the protein sequences that are to be
in the simulated run. Using one of 16 proteolytic enzymes and relevant digestion parameters
each protein sequence is in silico digested into peptides. Each peptide’s charge, mass, and
theoretical spectrum, including the isotopic distribution, is calculated. These calculations are
currently used to create centroided data. The simulator will be extended to create profile data in
the future. Centroided data will be most useful initially because most analytical software deals
with this type data. The simulator is implemented in the Ruby programing language and makes
use of and extends the mspire\textsuperscript{177} (mass spectrometry proteomics in Ruby) library. It is available
under the MIT license and works out of the box with sensible defaults. Customization to data
from different instruments is achieved through an included Ruby script, which uses a genetic
curve fitting algorithm. This script produces SVG files that visualize the fits as well as the necessary parameters for Mspire-Simulator to adapt its simulations.

The actual data used to create our default simulation model were obtained from our in-house LTQ-Orbitrap mass spectrometer coupled to reverse-phase liquid chromatography using nanospray ionization. The data are derived from an LCMS shotgun proteomic run of complex Human Embryonic Kidney (HEK-293T) cells. We used a Waters Nano Acuity column (15 cm long). ‘A’ solvent used was 95:5 water to acetonitrile and 0.1% formic acid and ‘B’ solvent was acetonitrile and 0.1% formic acid. Gradient was formed by 5–60% solvent mix over 70% of the run. These data along with all files produced and used are deposited at https://chorusproject.org/anonymous/download/experiment/17116340021687089 and at ProteomeXchange (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository\textsuperscript{286} with dataset identifier PXD000477. The MM14 data are already available at http://msbi.ipb-halle.de/msbi/centwave/, and the Orbitrap-Velos data are available upon request.

Orbitrap-Velos data were generously provided by the Christine Vogel lab and were from an ubiquitin pulldown from \textit{Saccharomyces cerevisiae} (Eksigent NanoFlow Plus, LC gradient 2–90% acetonitrile over 4.5 h at flow rate 400 nL/min). MM14 data are from the Bruker MicrOTOF-Q instrument and are described by Tautenhahn et al.; details can be found in that publication\textsuperscript{257}.

9.4 Results
Mspire-Simulator models elution, variance in intensity, and variance in the mass to charge ratio \((m/z)\) and predicts retention times and intensities for peptides. We follow the convention of Cappadona et al. and refer to a peptide feature as the full chromatographic profile of a peptide (at...
a given charge state) and a peptide peak as an individual isotopic component of a feature. Figure 9.1 outlines the overall process of simulation, and we consider each component in turn.

9.4.1 Retention Time and Intensity Prediction
A peptide is first situated along the retention time axis (See Figure 9.1A.). Both retention time and intensity are predicted for each peptide using a machine learned model built in WEKA. We used the M5Rules\textsuperscript{88} algorithm for retention time prediction and the M5P\textsuperscript{89} algorithm for intensity prediction, both of which gave the best correlation coefficients for our test data, \( \sim 0.96 \) and \( \sim 0.74 \), respectively, using the internal WEKA 10-fold cross validation technique. The test data can be downloaded as indicated in the methods section. These prediction models were trained on in-house data, which contained amino acid counts, average \( m/z \) value, the charge state, mass, retention time, and a binned intensity value for 1484 peptides. The intensity values were

![Figure 9.1: Overall process of simulation from theoretical spectrum to realistic peaks. The underscored 3d box in parts B and E designates the specific peak shown in the following parts. (A) Theoretical spectrum calculated for a certain peptide. (B) Ideal elution profiles are given to the spectrum. (C, D) Intensity variance is calculated for each peak in the elution profile. (E, F) Mass to charge variance is calculated for each peak in the elution profile.]

155
binned into 10 bins based on magnitude ranges. This allowed for a better prediction of intensities. The user may replace the default models with custom/better ones to mimic other configurations and instruments. For each peptide a single retention time and intensity is predicted; these values are used as starting points from which the retention times and intensities of all centroids related to a particular peptide are generated. The retention times are coerced into times that conform to a user-specified sampling rate (e.g., one scan per 2 s). Elution profiles are generated by sampling from the normal distributions of parameters $t$ and $f$.

9.4.2 Feature Shape

Peptide features are modeled along the m/z axis (Figure 9.1A) by predicting the charge states and isotope distribution of a peptide. For charge-state prediction, a user specifies a pH, after which a standard iterative procedure is used to determine the ratio of charge states that would be observed (e.g., for the peptide DRVYIHPF at a pH of 2.0, 29.045% of this peptide would have a charge of +2 and 70.959% a charge of +3). We label this parameter “ionized pH” to indicate that it represents the acidity of the peptide as it enters the mass spectrometer and not necessarily in LC buffer. Isotope distributions are calculated by FFT convolution.

The elution profile (Figure 9.1B) is produced by function composition of a dynamic standard deviation with a Gaussian function. The standard deviation ($\sigma$) is based on the relative position along the elution curve:

$$\sigma = tx + f$$

where $x$ is the relative retention time index from the starting retention time of the feature, $t$ is the tailing factor, and $f$ influences the shape at the front of the profile. The elution profile is then given by substitution of $\sigma$ into a Gaussian function:
where $i$ is the intensity at that point in the elution, $\mu$ is where the apex of the curve is located, $x$ is as above, and $h$ is an initial height factor, which determines the maximum height of the peak and is the same for each peptide. Thus, $i$ is a generalized intensity that is later modified by a variance model and predicted intensity values as previously mentioned. This produces a skewed elution profile that fits peaks derived from a wide variety of elution conditions, as shown for data an LTQ-Orbitrap (Figure 9.2A), a Quadrapole Time-of-Flight (Qq-TOF) (Figure 9.2B), and an Orbitrap-Velos (Figure 9.2C).

For the mass spectrometer types we examined there was a global relationship between the intensity of a peak and the variance of its measurement. We observed larger intensity variance in more intense features and thus also nearer the apex of an eluting peak (Figure 9.1C, D). An inverse exponential function captures this relationship:

$$\sigma = m \ast \left(1 - e^{-c \ast i}\right) + d$$

where $\sigma$ defines the standard deviation in intensity given the intensity value, $i$. The $c$, $d$, and $m$ parameters represent experimentally derived constants that can be used to fine-tune the function for different mass spectrometers or run conditions (Figure 9.3A, B). $\sigma$ is then composited into a
Gaussian function for each peak, again like above, and the ideal intensity is modified by drawing stochastically from this distribution. Our intensity variance model adequately mimics real data. When compared with actual data we observe an RMSD of 0.9051 (Figure 9.3A, B).

The $m/z$ variance is a function of intensity and therefore may vary between peptide features and also within each elution profile contained in a feature (Figure 9.1E, F). This is modeled by the following function:
\[ \sigma = m \cdot i^{-y} \]

where \( \sigma \) is the standard deviation, \( i \) is the relative intensity of the feature at that point in its elution profile, and \( y \) is another experimentally derived constant that can be fit to different data types. The standard deviation function is composited with a Gaussian function, similar to the above elution functions, and is then randomly sampled from to give the quantity of \( m/z \) variance in either direction. The \( m/z \) variance model produces realistic results based on the comparison of simulated \( m/z \) variance to actual \( m/z \) variance (Figure 9.3C, D).

Feature shape is further modeled by given protein abundances. The abundances can be specified in the FASTA file header by a “#” symbol, followed by a value representing the percentage of that protein in the sample. If no abundances are given, equal molarity is assumed. These values are then used to modify the total area under the function that determines feature shape by a simple scaling procedure.

9.4.3 Drops and Noise

At certain retention times, in real LC-MS runs, entire scans where very few if any peaks are observed are referred to as “drops” (e.g., PeptideAtlas accession PAe000142 contributed by S. Markey). Our model also simulates drops at random retention times by a specified percentage of the total run time. This elevates the realism in the simulation and adds another dimension of control when using simulated data to test analytical software.

The simulator has the ability to add white noise to the spectra based on density and intensity factors specified by the user. The higher the density factor, the more white noise there is in each spectrum. Intensities are pulled from a uniform distribution that varies between a maximum and
minimum value given by the user. These parameters, along with the option to turn off the white
noise completely, give the user complete control for testing purposes.

9.4.4 Merging Overlapping Peaks
As a final processing step, overlapping peaks are detected and merged. This is accomplished by
using a ppm range to define whether two peaks are sufficiently close to be overlapping or not.
The intensities of the peaks to be merged are summed, and the new m/z value is calculated by a
weighted average of the original m/z values weighted by the intensities of the respective peaks.
We use 1/4 of the m/z variance in ppm to define the range that we use to detect overlapping
peaks, and this parameter is adjustable by the user.

9.4.5 MS/MS
Theoretical fragmentation spectra are produced by generating fragment ion formulas for all
possible cleavages and calculating the mass for each ion at the predicted charge states. The ion
types are configurable, and masses can be average or monoisotopic. The fragmentation spectra
are produced by the MS-fragmenter gem, freely available from Rubygems.org.

9.4.6 Modifications
Mspire-Simulator has the ability to add modifications to specified residues and termini.
Modifications are read in by the user specifying a modification ID from the PSI-MOD.obo and
which residue/terminus to apply it too. These modifications are then used in the calculation of
each spectrum. Because there will always be modifications in peptide samples, this is an
important part of simulation.

9.4.7 Output
The simulated run is written to an mzML file that can be visualized with any mzML file viewer.
The mzML format is the standard de jure and is quickly becoming the standard de facto for mass
spectrometer data. Cross-platform converters like Proteowizard\textsuperscript{26} can convert mzML into mzXML\textsuperscript{291} or other formats. Alternatively, the code base itself could easily be extended to directly output other representations of the data as well. The program also creates a SQLite, XML, or CSV file, which contains information on all of the data in the simulated run, which can then be used to validate peak picking and quantitation software.

9.4.8 Parameter Fitting Automation
Simulating data from different mass-spectrometers and operating conditions requires some customization of noise and variance parameters. We developed a genetic algorithm to discover parameters from actual data. Figure 9.4 demonstrates the automatic fitting of Orbitrap Velos

\textbf{A.} normalized intensity vs retention time

\textbf{B.} m/z variance vs normalized intensity

\textbf{C.} intensity variance vs normalized intensity

\textbf{Figure 9.4: Visual output from the curve fitting algorithm. Max intensity normalization was used for each. This is a fit of Orbitrap Velos data. The blue dots show the actual data, and the red smooth lines represent the curve fit. This shows the ability to quickly generate parameters needed to simulate different types of data. This output took \(\sim 5\) min.}
data, and it works equally well on the many peaks and instrument types we have tested.

9.4.9 Using the Simulator to Assess Quantitation Performance
The lack of quantitative comparison of data processing and wet lab protocol is due in large part to the daunting task of obtaining labeled data.\textsuperscript{32,256} The size and complexity of MS data sets precludes obtaining labeled data without a significant outlay of resources. Mspire-Simulator provides a facile method for generating any quantity of labeled simulated data. As a case study, consider Smith et al., where Mspire-Simulator data used in conjunction with hand-labeled real data allowed the use of qualitative metrics to evaluate the accuracy of a peak summarization, a data-processing step in non-chromatographic studies.\textsuperscript{248}

![Figure 9.5: Comparison of simulated and actual MS features.](image)

Left side shows simulated features and right side shows the actual features. (A) Visual comparison of LC-MS feature from the peptide: HLVDEPQNLK (single-letter code amino acids). See Table 1 for analytical comparison. (B) Detail of a single elution profile showing m/z variance characteristics. Simulated m/z variance is very similar to actual (see Table 1; row 1−5).
Discussion and conclusion

Mspire-Simulator succeeds at creating highly realistic LC-MS peptide features, as demonstrated by the comparison between actual and simulated data shown in Figure 9.5. Under macro- and microscopic inspection, analytically and visually, the two features are virtually indistinguishable (Table 9.1). An entire simulated run of bovine serum albumin (BSA) (Figure 9.6) shows the similarity between the simulated data and what is commonly observed in performing an actual BSA digest during quality control runs. Mspire-Simulator can also produce highly complex runs.
(Figure 9.7) as well as simulate data from different mass spectrometers (Figure 9.2 and Figure 9.3).

With Mspire-Simulator’s abilities, layering simulated data next to or on top of actual data, visually or analytically, could become standard practice in proteomics, much the way MS/MS spectra are layered onto predicted b and y ion series to identify potential database matches. By comparing actual data with the model and then refining the model, a feedback loop is created that has utility not only in affirming what is known but in pointing out what is missing. Is an expected peptide missing because it has been modified? Are changes in the ratios of charge states indicative of pH or electrospray voltage aberrations? These and other aspects of a run can

![Figure 9.7: Bird’s-eye view of a simulated complex human cell run. 50 000 peptides were taken from the human FASTA database and simulated in two charge states creating 100 000 features. The run was generated in ~31 h on a single 2.50 GHz core and used ~1.9 GB of RAM. White vertical lines represent dropped/lower signal and are intentionally included. The run demonstrates the simulator’s ability to generate highly complex runs. Purple peaks are the highest intensity, then red, yellow, and gray is the lowest. Viewed in TOPPView.](image-url)
now be queried, and this process will inevitably result in more complete, more refined models of shotgun proteomics.

Refinements to Mspire-Simulator will focus initially on technical aspects of a LC-MS proteomic experiment. These include: a more explicit model of a peptide’s ionization efficiency;\textsuperscript{292,293} the pH of a solution as buffer concentrations change and as influenced by the electrospray process; the relative rates of tryptic digestion as a function of adjacent amino acid residues;\textsuperscript{294} profile data simulation; exploring the relationship between variance parameters and m/z and retention time; and improvements in peak merging.

Table 9.1: Statistics comparing the two features shown in Figure 9.5

<table>
<thead>
<tr>
<th>statistic</th>
<th>actual</th>
<th>simulated</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>m/z variance isotope index 1</td>
<td>0.095</td>
<td>0.215</td>
<td>0.120 ppm</td>
</tr>
<tr>
<td>(least abundant)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m/z variance isotope index 2</td>
<td>0.071</td>
<td>0.137</td>
<td>0.066 ppm</td>
</tr>
<tr>
<td>m/z variance isotope index 3</td>
<td>0.239</td>
<td>0.373</td>
<td>0.134 ppm</td>
</tr>
<tr>
<td>m/z variance isotope index 4</td>
<td>0.247</td>
<td>0.203</td>
<td>0.043 ppm</td>
</tr>
<tr>
<td>m/z variance isotope index 5</td>
<td>0.032</td>
<td>0.296</td>
<td>0.264 ppm</td>
</tr>
<tr>
<td>(most abundant)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intensity variance</td>
<td>26.3</td>
<td>25.8</td>
<td>0.52</td>
</tr>
<tr>
<td>elution time interval length (seconds)</td>
<td>43.385</td>
<td>43.360</td>
<td>0.025</td>
</tr>
<tr>
<td>normalized intensity mean</td>
<td>21.9</td>
<td>21.35</td>
<td>0.55</td>
</tr>
<tr>
<td>normalized intensity median</td>
<td>8.2</td>
<td>8.081</td>
<td>0.119</td>
</tr>
<tr>
<td>number of samples/peaks used</td>
<td>73</td>
<td>85</td>
<td>12</td>
</tr>
<tr>
<td>(peaks &gt;0.9 normalized abundance)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of peaks in quartile 1:</td>
<td>21</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>number of peaks in quartile 2:</td>
<td>27</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----</td>
<td>----</td>
<td>---</td>
</tr>
<tr>
<td>number of peaks in quartile 3:</td>
<td>15</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>number of peaks in quartile 4:</td>
<td>10</td>
<td>16</td>
<td>6</td>
</tr>
</tbody>
</table>

Mspire-Simulator could also be extended with more sophisticated modeling of biological phenomenon. More rigorous post-translational modification or splice-variant prediction would alter the landscape of predicted peptides. Protein-level enrichment could easily be added in, reflecting predictions about localization in a fractionated sample for instance. While biological questions are appropriately addressed after analysis of the raw data, it is nonetheless intriguing to consider mapping the biology as a simulated data set onto the raw data in an effort to generate putative identities for unanticipated peaks.

As simulated data becomes more sophisticated, we are aware of the possibility of its inappropriate use. The mzML file format is open and completely editable. As it currently stands, we see no way to prevent a simulated mzML file from being tampered with to be passed off as actual data, but the problem is not as hopeless as it might seem at first glance: the mzML format encourages the use of a file hash tag audit trail, so instrument produced data should always point back to a vendor produced raw data file. Deciding whether a file was simulated is roughly equivalent to deciding whether a file was tampered with and that is checking against a vendor-produced raw data file. The potential for the fraudulent use of simulated data should serve, then, to encourage what researchers should be doing anyway: providing access to raw data and using audit trails. In any event, we suggest that the potential benefits of simulation software far outweigh the challenges presented by potential misuse.

Mspire-Simulator will be useful initially in testing and developing algorithms for peak picking and quantitation. Simulated data are not meant to replace testing on actual data but to facilitate
more rigorous testing of algorithms. Data may be simulated with a range of peak and noise characteristics and strengths or flaws in algorithms uncovered. Mspire- Simulator will be especially useful in testing algorithms for their ability to accurately detect and quantify small peaks because the provenance of every centroid is known. Simulated data may ultimately facilitate workflows that find and quantitate an order of magnitude more peptides than is currently possible.
Chapter 10  Conclusion

This work describes efforts to improve upon the analysis methods and technologies available to identify and quantify biomolecules. As the challenges are diverse, we have introduced a complementary set of tools which have enabled novel discoveries, invited improvements to common practices, facilitated data quality, and provided novel identifications.

We have introduced a proposed nomenclature to clarify discussion of data analysis concepts in MS algorithm development. This project initiates a discourse about how to establish common ground and eliminate redundant efforts in algorithm development.

We have employed traditional mass spectrometry analysis and novel laboratory techniques to identify the protein components of a novel virus. We have elucidated complex structural motifs by MS analysis coupled with custom software implementations by both targeted MS3 and by consideration of statistical models for cyclized fragmentation.

Our independent implementation of the Kalman filter for peak detection outperforms competing algorithms by successfully identifying and quantifying isotope area with greater accuracy. This drastically improves accurate MS1 quantitation and feature finding. This work also reinforces the importance of the nomenclature introduced in Chapter 2 as necessary to efficiently express MS data analysis algorithms.

Lipid diversity complicates much of the identification and quantitation desired for lipidomic analyses. The most common approach to this complexity is segmentation of the problem space to provide reduced complexity to subsequent analysis. We introduced an automatic classification tool which is capable of classifying an unknown and uncharacterized lipid into reduced domain space for subsequent analysis. It is trained from the manually curated LMSD
and suggests improvements to that ontology, as well as providing an intelligible representation of classification logic in decision tree format.

We have introduced an algorithm to predict lipid fragmentation which is capable of predicting the complex cyclization products known to occur in lipids unlike existing solutions which fail to address the extremely diverse fragmentation schemes observed in lipid fragmentation studies. This work delivers a lipid fragmentation prediction software tool which follows published lipid fragmentation schemes to deliver a chemically-sound lipid fragmentation models which outperform existing solutions to identify selected lipids by their fragmentation spectra.

I have developed a software suite capable of automating performance metric generation designed to facilitate performance metric use by instrument operators and to improve general instrument performance. This suite archives performance metric and sample analysis runs and provides innovative graphical comparisons and statistical analysis in comparison to periodic control samples. The suite is web based and available on all major software platforms. This paper provides a tool and suggested MS omic workflow which incorporates automated and comprehensive quality control at the instrument stage.

We have provided the first LC/MS data simulation allowing for true ‘gold standard’ comparisons between analytical methods. The capabilities of this utility make it ideal for establishing an understanding of the relative strengths and weaknesses of proteomic MS data analysis.

Collectively, these projects have produced influential developments in the MS omic fields by encouraging quality control and rigorous comparisons among data analysis techniques, improving quantitation tools for label-free experiments, and introducing tools to identify complex peptide structural motifs and lipids. These developments have aided biomolecular...
understanding of neurotransmitter SNAP-25B structure, characterized a novel virus structure, and facilitated a variety of other experimental projects. The virtues of the techniques, algorithms, and tools developed will drive further biomolecular discoveries and innovations.
References


3. NIH IMPAC NIH Data Book - NIH Research Portfolio Online Reporting Tools (RePORT)

4. Nobel Media AB The Nobel Prize in Chemistry 2002


31. Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *Journal of proteome research* **2011**, *10*, 1785–93.


177


65. Tracy, C. M.; Gray, A. J.; Cuéllar, J.; Shaw, T. S.; Howlett, A. C.; Taylor, R. M.; Prince, J. T.; Ahn, N. G.; Valpuesta, J. M.; Willardson, B. M. Programmed Cell Death Protein 5 Interacts with the Cytosolic Chaperonin Containing Tailless Complex Polypeptide 1


73. Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M. Y.;
Vitek, O.; Aebersold, R.; Müller, M. SuperHirn - A Novel Tool for High Resolution LC-

74. Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized
P.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature

75. Conley, C. J.; Smith, R.; Torgrip, R. J. O.; Taylor, R. M.; Tautenhahn, R.; Prince, J. T.
Massifquant: Open-Source Kalman Filter Based XC-MS Isotope Trace Feature Detection.
*Bioinformatics (Oxford, England)* **2014**.

76. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.;
Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics
2010*, *26*, 966–968.

The Work of the Human Proteome Organisation’s Proteomics Standards Initiative (HUPO

of Terms Relating to Mass Spectrometry (IUPAC Recommendations 2013). *Pure and
Applied Chemistry* **2013**, *85*.

79. Wang, P.; Tang, H.; Fitzgibbon, M. P.; McIntosh, M.; Coram, M.; Zhang, H.; Yi, E.;
Aebersold, R. A Statistical Method for Chromatographic Alignment of LC-MS Data.


87. Zhang, X.; Asara, J. M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A. K. Data Pre-
Processing in Liquid Chromatography-Mass Spectrometry-Based Proteomics. 


101. Wommack, K. E.; Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems.

*Microbiology and Molecular Biology Reviews* 2000, 64, 69–+.  


*Extremophiles* 1997, 1, 143–149.  


104. Wikoff, W. R. Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid.


179. Bustos, C. Distribution.


216. ACD/Labs.com :: ACD/MS Fragmenter 


232. Hsu, F.-F.; Lobasso, S.; Turk, J.; Corcelli, A. Structural Studies on Archaeal Phytanyl-Ether Lipids Isolated from Membranes of Extreme Halophiles by Linear Ion-Trap


238. Hsu, F.-F.; Turk, J. Characterization of Cardiolipin from Escherichia Coli by Electrospray Ionization with Multiple Stage Quadrupole Ion-Trap Mass Spectrometric Analysis. 


239. Hsu, F.-F.; Turk, J. Characterization of Cardiolipin as the Sodiated Ions by Positive-Ion Electrospray Ionization with Multiple Stage Quadrupole Ion-Trap Mass Spectrometry. 


250. Taylor, R. M. (BYU); Miller, R. H. (BYU); Miller, R. D. (BYU); Porter, M. (BYU); Dalgleish, J. (BYU); Prince, J. T. (BYU) Automated Lipid Classification by Machine Learning. *Bioinformatics (Oxford, England)* **2014**.

251. Daylight Chemical Information Systems, I. Daylight>SMIRKS Tutorial


264. Welch, G.; Bishop, G. An Introduction to the Kalman Filter **1995**.


278. Highsoft Software Highcharts JS.


   Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.;
   Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R.
   A Common Open Representation of Mass Spectrometry Data and Its Application to

   Huang, S.-T.; Ferlanti, E. S.; Saeed, A. I.; Fleischmann, R. D.; Peterson, S. N.; Pieper, R.
   The APEX Quantitative Proteomics Tool: Generating Protein Quantitation Estimates from

293. Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martín, D.; Ranish, J.; Raught,
   B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational Prediction of