UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Generating functions of tandem mass spectra and their applications for peptide identifications

Permalink

https://escholarship.org/uc/item/7c97g96g

Author

Kim, Sangtae

Publication Date

2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Generating Functions of Tandem Mass Spectra and Their Applications for Peptide Identifications

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Sangtae Kim

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Vineet Bafna
Professor Nuno Bandeira
Professor Steven Briggs
Professor Charles Elkan

2012

Copyright
Sangtae Kim, 2012
All rights reserved.

The dissertation of Sangtae Kim is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2012

DEDICATION

To my beloved wife Jiyun, dearest son Eliah and daughter Chloe.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables x

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xv

Chapter 1 Introduction 1
1.1 The generating function approach 1
1.2 Evaluating statistical significance of Peptide-Spectrum Matches 2
1.3 Integrating de novo sequencing and database search . . . 3
1.4 Alternative to de novo sequencing: Spectral profiles and gapped peptides 3
1.5 Universal and sensitive database search tool 4

Chapter 2 The Generating Function Approach for Estimating Statistical Significance of Peptide-Spectrum Matches 6
2.1 Introduction 6
2.2 Methods 10
2.2.1 The generating function 10
2.2.2 Computing the generating function for boolean spectra 11
2.2.3 Computing the generating function for real spectra 12
2.3 Results 14
2.3.1 Datasets 14
2.3.2 Using generating functions to estimate the statistical significance of peptide identifications 15
2.3.3 Generating functions increase the number of identifications in MS/MS database searches 17
2.4 Discussion 20
2.5 Acknowledgements 21

Chapter 3 Integrating De Novo Sequencing with Database Search 30
3.1 Introduction 30
3.2 Methods 34
3.2.1 Peptide Sequencing Problem for Boolean Spectra 34
3.2.2 From Boolean spectra to MS/MS spectra 36
3.3 Results 43
3.3.1 Datasets 43
3.3.2 Generating multiple de novo reconstructions . . . 43
3.3.3 How existing database search approaches fare while searching very large databases? 44
3.3.4 Performance of MS-Dictionary 45
3.3.5 Using MS-Dictionary for database search 46
3.3.6 Searching the six-frame translation of human genome with MS-Dictionary 49
3.4 Discussion 52
3.5 Acknowledgements 53

Chapter 4 Spectral Profiles and Their Applications for de Novo Peptide Sequencing 67
4.1 Introduction 67
4.2 Methods 70
4.3 Results 74
4.4 Discussion 79
4.5 Acknowledgements 80

Chapter 5 Database Search of CID, ETD, and CID/ETD Pairs 92
5.1 Introduction 92
5.2 Methods 96
5.2.1 Digestion of cell lysate 96
5.2.2 Peptide pre-fractionation by strong cation exchange (SCX) 97
5.2.3 Mass spectrometry 97
5.2.4 Data processing 98
5.2.5 Mascot analysis 99
5.2.6 MS-GF training 99
5.2.7 MS-GFDB search (for CID or ETD spectra) . . . 102
5.2.8 MS-GFDB search (for CID/ETD pairs) 103
5.3 Results 104
5.3.1 Analysis of individual spectra 104
5.3.2 Comparison of ion fragmentation statistics across different spectral data sets 105
5.3.3 Pitfalls of “intersection” and “union” approaches to identifying CID/ETD pairs 106

5.3.4 Identifications from combined CID/ETD spectra . 107
5.4 Discussion 108
5.5 Acknowledgements 109

Chapter 6 Universal and Sensitive Database Search Tool 123
6.1 Introduction 123
6.2 Methods 126
6.2.1 Generating PRM spectra 126
6.2.2 Searching a protein database 129
6.2.3 Computing E-values 132
6.2.4 How to benefit from high-precision MS/MS spectra? 133
6.2.5 Estimating FDRs 138
6.3 Results 138
6.3.1 MS-GF+ Scoring 138
6.3.2 Datasets 140
6.3.3 Comparison of MS-GF+ and Mascot+Percolator . 143
6.3.4 Using MS-GF+ to identify peptides produced by αLP 144
6.3.5 Running time of MS-GF+ 145
6.3.6 Comparison of MS-GF+ with spectral library search 146
6.4 Discussion 147
6.5 Acknowledgements 148

Bibliography 161

LIST OF FIGURES

Figure 2.1: Illustration of the generating function 23
Figure 2.2: Illustration of the dynamic programming algorithm for computing the generating function 24
Figure 2.3: Separation between correct and incorrect identifications 25
Figure 2.4: Joint distribution of $SCORE$ and $Energy$ 26
Figure 2.5: Sensitivity-specificity trade-offs 27
Figure 2.6: Performance of MS-GF vs. X!Tandem 28
Figure 2.7: Accuracy of E-value estimates 29

Figure 3.1: Two approaches to peptide identification 56
Figure 3.2: Spectrum generation model and amino acid graph 57
Figure 3.3: Correlation between InsPecT and MS-Dictionary scores 57
Figure 3.4: Template-free recalibration 58
Figure 3.5: Correlation between peak intensity ranks and ion types 59
Figure 3.6: Example of two optimal de novo interpretations for a single spectrum 60
Figure 3.7: Example of a spectrum where the correct peptides gets a suboptimal de novo score 61
Figure 3.8: Fraction of spectra where the correct peptide has a suboptimal de novo score 62
Figure 3.9: MS-Dictionary accuracy as a function of the spectrum length . 63
Figure 3.10: Comparison of InsPecT, X!Tandem, MS-Dictionary, and MS-GF 64
Figure 3.11: Venn diagram showing the overlap between peptides identified by different approaches 65
Figure 3.12: Comparison of MS-Dictionary searches against the translated human genome and the human proteome 66

Figure 4.1: Example of a spectral profile 82
Figure 4.2: Various filtering approaches to peptide identifications 83
Figure 4.3: Overview of the MS-Profile tool 84
Figure 4.4: Example of the dynamic programming algorithm for computing the spectral profile 85
Figure 4.5: Distribution of spectral probabilities for PSM in the Standard dataset 86
Figure 4.6: Average accuracy and length of MS-Dictionary reconstructions and MS-Profile gapped peptides 87
Figure 4.7: Average accuracy and length of reconstructions generated by Peaks, PepNovo+, MS-Profile(Peaks) and MS-Profile(PepNovo+) 88
Figure 4.8: Average accuracy and length of multiple reconstructions of PepNovo+, MS-Profile(PepNovo+) 89

Figure 4.9: Comparison of accuracy of InsPecT tags and MS-Profile gapped peptides . . . 90
Figure 4.10: Accuracy of probabilities of profile peaks . . . 91

Figure 5.1: Computing p-values with MS-GF for a single spectrum . . . 111
Figure 5.2: Computing p-values with MS-GF for CID/ETD pairs . . . 112
Figure 5.3: Example of the offset frequency function (OFF) from the (charge-reduced) precursor m/z . . . 113
Figure 5.4: Number of identified peptides for Mascot and MS-GFDB . . . 114
Figure 5.5: Comparison of MS-GFDB with SEQUEST, OMSSA, and iProphet 115
Figure 5.6: Comparion of MS-GFDB and Percolator . . . 116
Figure 5.7: MS-GFDB is not susceptible to overfitting . . . 117
Figure 5.8: Probabilities of various ion types . . . 118
Figure 5.9: Rank distributions of different ion types for spectra of charge 2 119
Figure 5.10: Rank distributions of different ion types for spectra of charge 3 120
Figure 5.11: Venn diagrams of CID and ETD identifications . . . 121
Figure 5.12: Number of identified peptides with MS-GFDB CID/ETD . . . 122

Figure 6.1: Spectral types as paths in the graph . . . 149
Figure 6.2: Two approaches for searching a protein database . . . 150
Figure 6.3: Peptide generation model . . . 150
Figure 6.4: Illustration of the MS-GF+ scoring . . . 151
Figure 6.5: Illustration of the procedure constructing a G-spectrum . . . 152
Figure 6.6: Accuracy of the EFDR and the FDR via TDA . . . 153
Figure 6.7: Comparison of MS-GF+ and other tools for diverse spectral types 154
Figure 6.8: Comparison of MS-GF+ and others for phosphopeptides . . . 155
Figure 6.9: Comparison of MS-GF+ and others for αLP digests . . . 156
Figure 6.10: Running time of MS-GF+ and Mascot+Percolator . . . 157
Figure 6.11: Comparison of MS-GF+ and SpectraST . . . 158

LIST OF TABLES

Table 2.1: Prediction of decoy hits 22

Table 3.1: Comparison of MS-Dictionary and PepNovo 54
Table 3.2: Accuracy of InsPecT and X!Tandem for a search against all peptides 54
Table 3.3: MS-Dictionary identification of Shewanella spectra 55

Table 4.1: Accuracy and average length of PEAKS, PepNovo+ and MS-Dictionary for the Standard dataset 81

Table 6.1: Rescaling of amino acid masses 159
Table 6.2: Database search parameters used for MS-GF+ and Mascot+Percolator searches 160

ACKNOWLEDGEMENTS

First, I am grateful to my advisor, Prof. Pavel Pevzner. I first met Prof. Pevzner through his insightful books and papers, and had dreamed to work with him. The dream became reality in 2006, and I still have to pinch myself to believe that I am (and was when this is published) a student of Prof. Pevzner. He has been a great source of inspiration in my research and also in my life. I learned a lot from him, among which the most important may be how to enjoy research. He taught me this not through his words but through his smile during our discussions on various topics.

I am thankful to Prof. Nuno Bandeira for being a great mentor, colleague, friend, and also a member of my dissertation committee. I also thank Prof. Bafna, Prof. Briggs, and Prof. Elkan for serving in my dissertation committee.

All the former and current members in the Pevzner, Bafna, and Bandeira laboratories have been my constant source of inspiration and motivation. I am indebted to them, especially Julio Ng, Natalie Castellana, Ari Frank, Nitin Gupta, Samuel Payne, Xiaowen Liu, Jocelyne Bruand, Hosein Mohimani, Jian Wang, and Kyowon Jeong.

I am grateful to Prof. Eunok Paek and Prof. Heejin Park for introducing me to this wonderful area of mass spectrometry, and my former advisor Prof. Kunsoo Park for teaching me the beauty of computer algorithms. I am also grateful to my collaborators Drs. Richard Smith, Matthew Monroe, Albert Heck, and Elizabeth Komives for providing me with their valuable data and feedbacks on the software tools that I developed.

I am thankful to my parents and parents in law for supporting my study. In particular, without the devotion of my mother in law, I could never have completed this doctoral study. I thank my dearest kids Eliah and Chloe who have been infinite source of my energy - Daddy loves you so much.

Last not the least, I am invaluably grateful to my wife Jiyun who has been my biggest supporter throughout this long journey.

Chapter 2, in full, was published as “Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases”. S. Kim,

N. Gupta, and P. A. Pevzner. *Journal of Proteome Research*, vol. 7, no. 8, pp. 3354-3363, 2008. The dissertation author was the primary author of this paper.

Chapter 3, in full, was published as "Spectral Dictionaries: Integrating De Novo Peptide Sequencing with Database Search of Tandem Mass Spectra". S. Kim, N. Gupta, N. Bandeira, and P. A. Pevzner. *Molecular & Cellular Proteomics*, vol. 8, no. 1, pp. 53-69, 2009. The dissertation author was the primary author of this paper.

Chapter 4, in full, was published as "Spectral profiles: A Novel Representation of Tandem Mass Spectra and Its Applications for De Novo Peptide Sequencing and Identification". S. Kim, N. Bandeira, and P. Pevzner. *Molecular & Cellular Proteomics*, vol. 8, no. 6, pp. 1391-14009, 2009. The dissertation author was the primary author of this paper.

Chapter 5, in full, was published as "The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search". S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck, and P. Pevzner. *Molecular & Cellular Proteomics*, vol. 9, no. 12, pp. 2840-2852, 2010. The dissertation author was the primary author of this paper.

Chapter 6 is in preparation for publication as "MS-GF+: Universal and Sensitive Database Search Tool for Mass Spectrometry". S. Kim, and Pavel Pevzner, in preparation. The dissertation author is the primary author of this paper.

VITA

2000	Bachelor of Science in Computer Engineering, Seoul National University, Seoul, Korea
2002	Master of Science in Computer Science, Seoul National University, Seoul, Korea
2012	Doctor of Philosophy in Computer Science, University of California, San Diego

PUBLICATIONS

Sangtae Kim, and Pavel Pevzner, MS-GF+: A Sensitive and Adaptive Database Search Engine for Mass Spectrometry, submitted.

Kyowon Jeong, Sangtae Kim, and Nuno Bandeira, False Discovery Rates in Spectral Identification: a Target-Decoy Approach, submitted.

To-ju Huang, Claudiu Farcas, Jeremy Carver, Natalie Castellana, Ari Frank, Sangtae Kim, Jian Wang, Xiaowen Liu, Pavel A. Pevzner, Vineet Bafna, Ingolf Krüger, and Nuno Bandeira, ProteoSAFe: A Scalable, Accessible, and Flexible Software Environment for Proteomics Analysis, in preparation.

Kyowon Jeong, Sangtae Kim, Nuno Bandeira, and Pavel Pevzner, Gapped Spectral Dictionaries and Their Applications for Database Searches of Tandem Mass Spectra, Molecular & Cellular Proteomics, 10, M110.002220, 2011.

Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J. Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J. R. Heck, and Pavel Pevzner, The Generating Function of CID, ETD and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search, Molecular & Cellular Proteomics, 9, 2840-2852, 2010.

Kyowon Jeong, Sangtae Kim, Nuno Bandeira, and Pavel Pevzner, Gapped Spectral Dictionaries and Their Applications for Database Searches of Tandem Mass Spectra, In Proceedings of the Fourteenth International Conference on Research in Computational Molecular Biology (RECOMB-2010), 208-232, 2010.

Sangtae Kim, Nuno Bandeira, and Pavel Pevzner, Spectral Profiles: A Novel Representation of Tandem Mass Spectra and its Applications for de Novo Peptide Sequencing and Identification, Molecular & Cellular Proteomics, 8, 1391-1400, 2009.

Sangtae Kim, Nitin Gupta, Nuno Bandeira and Pavel Pevzner, Spectral Dictionaries: Integrating De Novo Peptide Sequencing with Database Search of Tandem Mass Spectra, Molecular & Cellular Proteomics, 8, 53-69, 2009.

Pavel Pevzner, Sangtae Kim, and Julio Ng, Comment on Protein Sequences from Mastodon and *T*yrannosaurus Rex Revealed by Mass Spectrometry, Science, 321 (5892), 1040, 2008.

Sangtae Kim, Nitin Gupta, and Pavel Pevzner, Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases, Journal of Proteome Research, 7, 3354-3363, 2008.

Sangtae Kim, Seungjin Na, Ji Woong Sim, Heejin Park, Jaeho Jeong, Hokeun Kim, Younghwan Seo, Jawon Seo, Kong-Joo Lee, and Eunok Paek, MODi : A Powerful and Convenient Web Server for Identifying Multiple Post-translational Peptide Modifications from Tandem Mass Spectra, Nucleic Acids Research, 34, W258-W263, 2006.

Sangtae Kim, Jeong Seop Sim, Heejin Park, Kunsoo Park, Hyunseok Park, and Jeong-Sun Seo, A Heuristic Algorithm to Find All Normalized Local Alignments Above Threshold, Genomics & Informatics, 1, 26-32, 2003.

ABSTRACT OF THE DISSERTATION

Generating Functions of Tandem Mass Spectra and Their Applications for Peptide Identifications

by

Sangtae Kim

Doctor of Philosophy in Computer Science

University of California, San Diego, 2012

Professor Pavel A. Pevzner, Chair

Mass spectrometry (MS) has become the leading high-throughput technology for proteomics, a large-scale study of proteins. MS experiments generate tandem mass (MS/MS) spectra, each representing a peptide. Identifying peptides from MS/MS spectra is a basic and essential task in proteomics studies. At present, MS instruments and experimental protocols are rapidly advancing, however, the software tools to interpret MS/MS spectra are lagging behind with many computational problems remaining unsolved. In this dissertation, we present a novel approach to interpreting MS/MS spectra, called the generating function approach, and show how this approach enables us to solve key computational problems in MS. First, we address the problem of estimating statistical significance of Peptide-

Spectrum Matches (PSMs). Since typically less than 30% of the generated spectra can be correctly interpreted, this problem is important in distinguishing between correct and incorrect PSMs. Using the generating function approach, we present the first analytical (rather than empirical) solution to this problem. Our MS-GF tool not only improves the accuracy of statistical significance estimates, but also increases the number of peptide identifications at a fixed error rate. Next, we present an alternative approach to peptide identifications based on generating all plausible de novo interpretations of a spectrum (spectral dictionary) and then quickly matching them against the protein database. Our MS-Dictionary tool enables proteogenomic searches in six-frame translation of genomic sequences that may be prohibitively time-consuming with traditional methods. We also present spectral profiles, a new representation of tandem mass spectra that compactly represent spectral dictionaries. Spectral profiles can be used to generate gapped peptides that are as useful as full-length peptides and as accurate as peptide sequence tags of length 3 traditionally used to speed up database searches. Lastly, we present a new database search tool MS-GF+ based on MS-GF. MS-GF+ is sensitive (it identifies more peptides than other database tools) and universal (works well for diverse types of spectra, different configurations of MS instruments and different experimental protocols). We benchmark MS-GF+ using diverse types of spectral datasets, and show that for all these datasets, MS-GF+ significantly increased the number of identified peptides compared to state-of-the-art methods for peptide identifications.

Chapter 1

Introduction

Mass spectrometry (MS) has become the leading high-throughput technology for proteomics [1]. MS experiments generate tandem mass (MS/MS) spectra, each representing a peptide. Identifying peptides from MS/MS spectra is a basic and essential task for most proteomics studies. However, this task remains challenging because the spectra are complex and the process of peptide fragmentation is not well understood. This dissertation presents a novel approach to interpreting MS/MS spectra, called the generating function approach, and describes its applications for peptide identifications.

1.1 The generating function approach

To introduce the notion of the generating function of tandem mass spectra, we use the analogy with the classical *Ising model* of ferromagnetism, one of the pillars of statistical mechanics [2]. The model consists of n magnetic spins such that each spin can be in two states (up and down). This results in 2^n possible *states* each with its own *energy* defined by the elementary interactions between neighboring spins on the lattice. The *partition function* represents the key technique for analyzing the Ising model and is defined as $\sum_{all\ states\ \pi} e^{-Energy(\pi)}$ (here we ignore the "temperature" parameter of the Ising model).[1]

[1]Partition functions in statistical mechanics represent a special class of generating functions and we use them below only to illustrate this notion in application to tandem mass spectra.

Interpreting a spectrum S with a peptide P is not unlike choosing a state in the Ising model. Instead of 2^n states of magnetic spins, there are 20^n possible *interpretations* of the spectrum S by peptides of length n. Each of these interpretations has its own "energy" given by the score of the match between spectrum S and peptide P. The goal is to compute the partition (generating) function of the spectrum S and to apply it to analyzing statistics of the MS/MS searches rather than the statistics of the Ising model. While the generating function of tandem mass spectra involves 20^n terms, we show in Chapter 2 how to efficiently compute it.

Chapter 2 introduces the notion of the generating function for MS/MS spectra using the Boolean spectrum that ignores intensities, charges, inaccuracies in peak positions, and different types of ions. While Boolean spectra are impractical, they proved to be useful as a stepping stone for introducing simple scoring/algorithms. Later in Chapter 2, we illustrate how to define the generating function for real spectra.

1.2 Evaluating statistical significance of Peptide-Spectrum Matches

A key problem in computational proteomics is distinguishing between correct and incorrect peptide identifications. Chapter 2 describes how the generating functions of MS/MS spectra and their derivatives (spectral energy and spectral probability) represent new features of tandem mass spectra that, similar to the commonly used Δ-scores, significantly improve peptide identifications. Furthermore, the spectral probability provides a rigorous solution to the problem of computing statistical significance of Peptide-Spectrum Matches (PSMs). Our MS-GF tool reports spectral probabilities of PSMs, improving the sensitivity-specificity trade-off of existing MS/MS search tools. It also addresses the notoriously difficult problem of one-hit-wonders in MS, and often eliminates the need for decoy database searches, a common approach used to control false positive peptide identifications. We show the generating function approach has the potential to increase the num-

ber of peptide identifications in MS/MS searches while effectively controlling false positives.

1.3 Integrating de novo sequencing and database search

There are two major approaches to identifying peptides from MS/MS spectra: De novo sequencing and database search. The de novo sequencing approach directly infers peptides from spectra, whereas the database search approach compares and scores spectra against theoretical spectra predicted from peptides in a protein database, and selects the best-scoring peptide. While de novo sequencing is emerging as an alternative to database search, database search remains a more accurate and thus preferred method. We studied an alternative approach where *all* plausible de novo interpretations of a spectrum (*spectral dictionary*) are generated and then quickly matched against the database [3]. Chapter 3 presents a new MS-Dictionary algorithm for efficiently generating spectral dictionaries using generating functions. We demonstrate that MS-Dictionary can identify spectra that are missed in the database search. MS-Dictionary enables proteogenomic searches [4] against six-frame translation of genomic sequences that may be prohibitively time-consuming for existing database search approaches. We show that such searches allow one to correct sequencing errors and find programmed frameshifts.

1.4 Alternative to de novo sequencing: Spectral profiles and gapped peptides

Despite many efforts in the last decade, the progress in de novo peptide sequencing has been slow with only 30–45% of all peptides being correctly reconstructed. In Chapter 4, We argue that accurate *full-length* peptide sequencing may be an unattainable goal for some spectra and demonstrate how to accurately sequence *gapped* peptides instead using the generating function approach. Gapped

peptides are nearly as useful as full-length peptides for error-tolerant database searches, occupying a niche between long but inaccurate full-length reconstructions and short but accurate peptide sequence tags. Gapped peptides are longer and more accurate than peptide sequence tags of length 3 traditionally used to speed up database searches in proteomics. To generate gapped peptides, our MS-Profile tool uses spectral profiles, a new representation of tandem mass spectra. Spectral profiles also enable intuitive visualization of all high scoring de novo reconstructions of tandem mass spectra.

1.5 Universal and sensitive database search tool

MS instruments and experimental protocols are rapidly advancing, but the software tools to analyze MS/MS spectra are lagging behind. While existing database search tools perform well on some types of spectra (e.g., Collision Induced Dissociation (CID) spectra of tryptic peptides), their performance often deteriorates on other types of spectra, such as Electron Transfer Dissociation (ETD), Higher-energy Collisional Dissociation (HCD) spectra, or spectra of non-tryptic digests.

Chapter 5 describes ideas on how to make MS-GF adaptable to *any* type of spectra and presents a new database search tool MS-GFDB based on MS-GF. MS-GFDB greatly outperforms a popular database search tool Mascot [5] for ETD spectra. Moreover, even after a decade of Mascot developments for analyzing CID spectra of tryptic peptides, MS-GFDB (that is not particularly tailored for CID spectra or tryptic peptides) resulted in significant increase over Mascot in the number of peptide identifications. We also propose a statistical framework for analyzing multiple spectra known to be generated from the same peptide (e.g. CID/ETD spectral pairs) and assigning P-values to Peptide-Spectrum-Spectrum Matches (PS^2Ms).

Although MS-GFDB showed promising results for interpreting various types of spectra, it has several limitations, such as the limited support of search for modified peptides and inability to benefit from high-precision MS/MS spectra.

Chapter 6 presents a new database search tool MS-GF+ that addresses all these limitations, greatly reduces the running time, and features a greatly improved usability. MS-GF+ is sensitive (it identifies more peptides than other database tools) and universal (works well for diverse types of spectra, different configurations of MS instruments and different experimental protocols).

We benchmark MS-GF+ using diverse spectral datasets: (i) spectra of varying fragmentation methods with either linear ion trap or orbitrap readout; (ii) spectra of multiple enzyme digests; (iii) spectra of phosphorylated peptides; (iv) spectra of peptides with unusual fragmentation propensities produced by a novel alpha-lytic protease. For all these datasets, MS-GF+ significantly increased the number of identified peptides compared to state-of-the-art methods for peptide identifications. We emphasize that while MS-GF+ is not specifically designed for any particular experimental set-up, it improves upon the performance of tools specifically designed for these applications. We also compare MS-GF+ with a leading spectral library search tool, SpectraST [6]. The spectral library search is a new emerging approach to use previously identified spectra for peptide identifications. While the existing view is that spectral library searches greatly improve on database searches (for previously identified peptides) [7], we show that MS-GF+ identifies a similar number of peptides as compared to SpectraST without using spectra in the library.

Chapter 2

The Generating Function Approach for Estimating Statistical Significance of Peptide-Spectrum Matches

2.1 Introduction

MS experiments often generate millions of spectra, and interpreting them leads to challenging statistical problems (see Nesvizhskii et al., 2007 [8] and Kall et al., 2008 [8] for recent reviews). One of the major problems in tandem mass spectrometry is the lack of theoretical (as opposed to empirical) estimates of statistical significance of peptide identifications. Indeed, the Proteomics Publication Guidelines [9, 10] recommend searching in decoy databases to determine the statistical significance of peptide identifications (this is in contrast to genomics searches that do not employ decoy databases). We argue that if the error rates reported by existing MS/MS software tools were reliable (as in the case of genomics searches), the search in decoy databases would not be necessary. The major difference here is that MS/MS searches are currently based on empirical database-dependent estimates of error rates (often represented by Poisson, Gaussian, hy-

pergeometric, or other approximations of tails of score distributions [11, 12, 13]) as opposed to the analytically derived and database-independent error rates in genomics tools like BLAST [14]. Although the target-decoy search strategy is currently viewed as the best way to distinguish between the correct and false identifications [15, 16, 17, 18, 19, 20], this valuable approach has certain shortcomings. While the shortcomings of such strategies are well recognized in genomics (see [21]), they are often overlooked in proteomics. Also, decoy databases take a toll on every lab engaged in MS/MS searches effectively doubling the search time. We argue that using decoy databases is an acknowledgment of our inability to solve the following problem:

Spectrum Matching Problem. Given a spectrum S and a score threshold T for a spectrum-peptide scoring function, find the probability that a random peptide matches the spectrum S with score equal to or larger than T.

The Spectrum Matching Problem was first posed by Fenyo and Beavis, 2003 [22] (see also [23]).[1] They acknowledged that the theoretical solution of this problem is unknown and suggested a heuristic approach to its solution based on approximating the tail of the score distribution. Solving the Spectrum Matching Problem is equivalent to computing the FPRs of spectral matches. FPR is a property of an *individual* spectrum as opposed to the False Discovery Rate (FDR), the property of *multiple* spectra (proportion of incorrect identifications among all identifications judged correct).[2]

Search in a decoy database looks like an attractive approach for approximating the solution of the Spectrum Matching Problem as $\frac{m}{n}$, where m is the number of matches between the spectrum and the decoy database of size n (with scores equal to or larger than the threshold T). However, for an *individual* spectrum, the number of matches for typical n is usually zero thus making this approach

[1]The Spectrum Matching Problem assumes a certain probabilistic distribution on the set of all peptides and computes the total probability of all peptides P with $score(P, S) > T$.

[2]Different papers on statistics of MS/MS searches often use inconsistent terminology. The solution of the Spectral Matching Problem provides E-values (the expected number of peptides with the scores equal to or larger than the observed score) or can be used for computing p-values in the hypothesis testing framework. To avoid a confusion, we follow the terminology from the recent review [8] and use the term FPR (and the related term *Spectral Probability* defined below) in the remainder of this chapter.

problematic (decoy and target databases usually have the same size). To obtain reliable FPR for an individual spectrum, one can increase n (e.g., making giant decoy databases 1000 times larger than target databases). Since this is impractical, some existing approaches bundle all spectra with the same score to evaluate the FDR of all spectra in the bundle and to use FDR as a surrogate for FPR (see [8]).

Assigning the same FPR to all identifications with identical scores [24, 5, 25] is a dangerous oversimplification since the scoring functions of existing MS/MS tools are not based on rigorous probabilistic models and are often inaccurate. Recognizing this problem, Fenyo and Beavis, 2003 [22] pioneered computing FPR for an *individual spectrum* as an empirical solution of the Spectrum Matching Problem.[3] They constructed the empirical score distribution of low-scoring (erroneous) peptide identifications and extrapolated it to evaluate the FPR of high-scoring peptide identifications in the tail of the distribution. Such approaches are not free of shortcomings: Waterman and Vingron, 1994 [21] wrote: "Theory is needed because simulations rarely cover the extreme tails of a distribution." criticizing similar approaches in genomics. In another paper criticizing such empirical approaches, Nagarajan et al., 2005 [26] demonstrated that *all* existing motif finding tools are statistically flawed and can be off by orders of magnitude in computing FPRs. This flaw remained uncovered for 15 years and affected 1000s of studies. Needless to say, the mass spectrometry community is not immune to similar flaws suggesting that re-examination of existing approaches to estimation of statistical significance in MS/MS searches is timely. In this chapter, we demonstrate that the analysis of statistical significance in various MS/MS tools is often unreliable (Figure 2.7).

We further argue that use of decoy databases is not free from shortcomings. The intuition behind using a decoy database is to estimate the number of spectra that match the database by chance. If a spectrum S has probability $p(S)$ of matching a random database, then a decoy database is simply a time-consuming way to evaluate $\sum p(S)$ over *all spectra* in the dataset (this sum represents the expected

[3]The approach in [22] is particularly attractive since it can be implemented without decoy databases.

number of hits in the decoy database) but not a good way to estimate individual probabilities $p(S)$. The generating function approach, in difference from the decoy database approach, accurately computes probabilities $p(S)$ for the individual spectra, an important advantage for addressing the problem of "one-hit-wonders" in MS/MS searches. An ideal approach to evaluating the statistical significance of MS/MS searches would be to use a database containing all possible peptides up to a certain length, and use the number of identifications in this database to evaluate the error rate. However, the time required to search this database renders this approach infeasible. Below we show that it is nevertheless possible to compute the precise number of the identified peptides in this huge database thus computing the solution of the Spectrum Matching Problem exactly rather than empirically. This illustrates the advantages of (fast) analysis of scores over the huge database of all peptides as compared to (slow) analysis of scores over the much smaller decoy databases.

Solving the Spectrum Matching Problem is not unlike computing the *generating function* in combinatorics [27, 28]. Given a spectrum S and a score X, define $E(S, X)$ as the number of peptides (among all possible peptides) that match the spectrum S with score X. To evaluate FPRs one has to compute $E(S, X)$ for every spectrum S and every score X (more precisely, the sum of probabilities of all peptides contributing to $E(S, X)$). Figure 2.1(b) illustrates the notion of the generating function in the simple case when the score X of a match between a spectrum and a peptide is defined as the number of peaks in the spectrum explained as b or y ions. Figure 2.1(c) shows the generating function for a more advanced scoring described below. We show how to compute $E(S, X)$ and to use it for improving the sensitivity-specificity trade-off of various database search tools. We further introduce the notion of *spectral energy* (Figure 2.1) that represents the difference between the best de novo spectral interpretation and the best database spectral interpretation. We show that while the *Energy*-score (in difference from the Δ-score) was ignored in MS/MS searches so far, it greatly improves the separation between the correct and false identifications. Finally, we introduce the notion of *spectral probability* (the total probability of all peptides with scores exceeding

a threshold) that further improves the separation between the correct and false identifications (Figure 2.1).

While this chapter is limited to identifications of non-modified peptides, the generating function approach can be extended to modified peptides as well (see Chapter 6). Our MS-GF software for computing generating function/spectral energy/spectral probability of tandem mass spectra is available as open source from http://proteomics.ucsd.edu/Software.html.

2.2 Methods

2.2.1 The generating function

For the sake of simplicity, we first introduce the notion of generating function for *boolean* spectra that ignore intensities, charges, inaccuracies in peak positions, and C-terminal ions. While the boolean spectra are impractical, they proved to be useful as a stepping stone for introducing simple scoring/algorithms and later generalizing them to real spectra and more complex algorithms (see [29, 30, 31]). Later, we will illustrate how to define the generating function for real spectra.

We represent a boolean spectrum S with parent mass k as 0-1 vector $s_1 \dots s_k$, where $s_i = 1$ if there is a peak at mass i in the spectrum, and $s_i = 0$, otherwise. This representation assumes that the spectra are discretized and all masses are integers (Figure 4.4). For example, for ion-trap spectra this can be approximated by multiplying all masses by 10 and taking integer parts (see Chapter 3 for details). The *match score* between spectra $s_1 \dots s_k$ and $s'_1 \dots s'_k$ is defined as $\sum_{i=1}^{k} s_i \cdot s'_i$.

Given a peptide $P = p_1 \dots p_n$, we define its theoretical spectrum $Spectrum(P)$ as a 0-1 spectrum $s_1 \dots s_k$ with $(n-1)$ 1s, such that $s_i = 1$ iff i is the mass of the peptide $p_1 \dots p_i$. The score (denoted as $Score(P, S)$) between a peptide P and a spectrum S (with the same parent mass) is defined as the match score between spectra $Spectrum(P)$ and S. For convenience, we assume that $Score(P, S) = -\infty$ if peptide P and spectrum S have different parent masses. Let $SCORE = SCORE(S) = \max_{all\ peptides\ P} Score(P, S)$ be the maximum value of

$Score(P, S)$ among all possible peptides P. $SCORE$ can be estimated using de novo peptide sequencing algorithms [32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43]. We define *energy* of a peptide-spectrum pair as $Energy(P, S) = SCORE - Score(P, S)$ and define the *generating function* of the spectrum S as $\sum_{all\ peptides\ P} e^{-Energy(P,S)} = \sum_t x(t) \cdot e^{-t}$, where $x(t)$ is the number of peptides with energy t.[4]

Given the probabilities of individual amino acids (e.g., computed empirically from a set of protein sequences), we define the probability $prob(P)$ of a peptide $P = a_1 \dots a_m$ as the product of probabilities of its amino acids $\prod_{i=1}^{m} prob(a_i)$. We will also consider the *weighted generating function*: $\sum_{all\ peptides\ P} prob(P) \cdot e^{-Energy(P,S)} = \sum_t y(t) \cdot e^{-t}$, where $y(t)$ is the overall probability of all peptides with energy t.

2.2.2 Computing the generating function for boolean spectra

Given a spectrum S, we introduce a variable $x(i, t)$ equal to the number of peptides of mass i that have t peaks in common with spectrum S, i.e, the number of peptides P such that $Score(P, S_i) = t$ (S_i stands for "i-prefix" $s_1 \dots s_i$ of the spectrum S). In the case S has a peak at position i ($s_i = 1$), the variable $x(i, t)$ can be computed as follows ($|a|$ denotes the mass of an amino acid a):

$$x(i, t) = \sum_{\text{all amino acids } a} x(i - |a|, t - 1)$$

Otherwise ($s_i = 0$):

$$x(i, t) = \sum_{\text{all amino acids } a} x(i - |a|, t)$$

[4]This expression represents the *exponential generating function* [27] of the vector $x = (x(0), x(1), \dots)$. Similarly to many applications of generating functions outside physics, we follow Herbert Wilf's interpretation of generating functions ("a clothesline on which we hang up a sequence of numbers" as defined in [28]) rather than using it as a model of a physical process. As some other applications of generating functions in bioinformatics [44], we do not analyze the analytical behavior of the MS/MS generating functions in this chapter.

Below we provide an equivalent and more compact representation of these recurrencies:

$$x(i,t) = \sum_{\text{all amino acids } a} x(i - |a|, t - s_i)$$

We initialize $x(0,0) = 1$, $x(0,t) = 0$ for $t > 0$, and assume that $x(i,t) = 0$ for negative i. The maximum value $SCORE$ of $Score(P,S)$ among all possible peptides P is simply the maximum value of t with non-zero $x(k,t)$. See Figure 4.4.

The recurrence for computing the weighted generating function is very similar. In this case the variable $y(i,t)$ equals to the overall probability of peptides of mass i that have t peaks in common with spectrum S. The variable $y(i,t)$ is initialized in the same way as $x(i,t)$[5] and is computed using the following recurrence:

$$y(i,t) = \sum_{\text{all amino acids } a} y(i - |a|, t - s_i) \cdot prob(a)$$

The above algorithm for computing the generating function has complexity $O(|S| \cdot |SCORE| \cdot Mult \cdot PeptideLength \cdot A)$, where $A = 20$ is the number of amino acids, $PeptideLength$ is the maximum length of a peptide with the mass equal to $|S|$, and $Mult$ is the multiplication coefficient that was applied to all masses in the spectrum to satisfy the assumption that they are integers (typically, $Mult = 10$ for ion-traps). In practice, it requires 0.1-0.2 seconds to compute the generating function on a desktop machine with 2.16 Ghz Intel processor.

2.2.3 Computing the generating function for real spectra

MS-GF transforms tandem mass spectra into its integer-valued scored version $s_1 \dots s_k$ (rather than boolean spectra) using the probabilistic model similar to [33, 35, 43]. This transformation takes into account peptide length, peak intensities, neutral losses, dependencies between ion types, noise, etc. Most de novo and database search algorithms use such representation (explicitly or implicitly) by assigning intensity-dependent scores to peaks, further adjusting for imprecisions in mass-measurements, and applying dot-product for scoring spectra against

[5] We initialize $x(0,0) = 1$ since the "empty" peptide is the only peptide with mass 0 that has 0 peaks in common with the spectrum S. We initialize $y(0,0) = 1$ since the probability of the empty peptide is defined as 1.

peptides. However, these scores are typically attached to the positions of peaks in the spectrum $s_1 \dots s_k$ and will not enable a computation of the generating function in the low-accuracy setting with accuracy threshold δ. However, as long as we redefine the spectrum $s_1 \dots s_k$ as $s'_1 \dots s'_k$ with $s'_i = \max_{j=i-\delta}^{j=i+\delta} s_j$, the generating function (in case of imprecise mass measurements) can be easily computed as described below.

The score $Score(P, S)$ between a peptide P and a spectrum S (with the same parent mass) is defined as the dot-product between the theoretical spectrum $Spectrum(P)$ and S (now S is defined as an arbitrary integer-valued vector and $Spectrum(P)$ is defined to allow for both N-terminal and C-terminal ions as in [31]). Let $SCORE$ be the maximum value of $Score(P, S)$ and $Energy(P, S) = SCORE - Score(P, S)$. Given a spectrum S, we define $x(i, t)$ as the number of peptides of mass i with score t, i.e, the number of peptides P such that $Score(P, S_i) = t$. The variable $x(i, t)$ can be computed as in the case of boolean spectra.

We emphasize that MS-GF can handle scored spectra generated by any MS/MS tool with additive scoring functions. The scoring function chosen here can be viewed as a variation of Sherenga and PepNovo [33, 35] with improved analysis of peak intensities and doubly charged ions (the details are described in Chapter 3). Some MS/MS analysis tools (e.g., SEQUEST or tools using sequence-specific peak intensities [45, 46, 47]) have non-additive scoring components and thus cannot be modeled by this generating function framework. However, MS-GF still can be used to re-score their results (Such re-scoring usually improves on non-additive scoring).

Let $\mathcal{A}$ be a peptide identification algorithm that accepts a peptide P as an interpretation of a spectrum S as long as the peptide-spectrum score $Score(P, S)$ is larger or equal to the threshold T. Given the allowed (integer) parent mass error ϵ, the weighted generating function allows one to compute the overall probability of peptides with scores equal to or larger than T (*spectral probability*) as

$$Prob_T(S) = \sum_{i=ParentMass-\epsilon}^{i=ParentMass+\epsilon} \sum_{t \geq T} y(i, t)$$

For example, the spectral probability $Prob_{60}(S) = 2.76 \cdot 10^{-10}$ represents the

total probability of all 306 peptides with scores larger or equal to the score of the correct peptide in Figure 2.1(c). The probability that the algorithm $\mathcal{A}$ identifies the spectrum S in a random database of size n is computed as $1-(1-Prob_T(S))^n$. Since the parameter T is usually chosen in such a way that $Prob_T(S)$ is much smaller than $\frac{1}{n}$, one can assume that $1-(1-Prob_T(S))^n \approx Prob_T(S)\cdot n$. If a user attempts to identify peptides with a fixed FPR in a database of size n (e.g., $FPR = 0.01$ is commonly used in MS/MS searches), then the parameter T is chosen in such a way that $Prob_T(S) = \frac{FPR}{n}$. The corresponding value of T can be derived from the generating function (see the last column in Figure 2.1(c)).

2.3 Results

2.3.1 Datasets

The *Shewanella oneidensis* MR-1 dataset used here (14.5 million spectra) and peptide identifications based on this dataset are described in [48]. 28,377 unmodified peptides were identified in this dataset by InsPecT with an error rate of 5% (1% spectrum-level error rate) as measured using a decoy database [25].

Due to its large size, searching the entire *Shewanella* dataset with tools like SEQUEST is rather time-consuming. To make it easier to benchmark our approach against other tools and to summarize the results, we constructed two smaller datasets (geared to peptides of length 10) that are used in this study. The results are similar for other peptide lengths.

- *Shewanella-1784*: From 28,377 peptides identified in *Shewanella oneidensis* MR-1, we selected all doubly-charged tryptic peptides of length 10. It resulted in 1745 and 39 peptides identified in the target and decoy databases (2.2% error rate). For each of these $1745 + 39 = 1784$ peptides, we retained one spectrum (chosen randomly if the peptide is identified from multiple spectra) to construct the final dataset of 1784 spectra.

- *Shewanella-50000*: From all 14.5 million *Shewanella* spectra, we randomly selected 50,000 doubly-charged spectra with parent masses ranging from 1100

to 1200 Da (these spectra typically correspond to peptides of length $\approx$ 10 aa). Each spectrum in this dataset was searched against all Shewanel la proteins (1.47 million of amino acids) and against the randomized decoy database (of same size) with SEQUEST (TurboSEQUEST v.27, rev. 12), InsPecT (20060907), and X!Tandem (2007.01.01.2), as well as analyzed with MS-GF and PeptideProphet (v3.0).

2.3.2 Using generating functions to estimate the statistical significance of peptide identifications

We found that the error rates reported by existing database search tools do not provide accurate estimates of the statistical significance of *individual* peptide identifications (they are often off by an order of magnitude) while the error rates evaluated by MS-GF are very accurate (see Figure 2.7).

To evaluate whether MS-GF accurately estimates the number of hits in decoy database (thus eliminating the need for the decoy database search) we conducted the following experiment. For each spectrum in the *Shewanella-50000* dataset, we generated top-scoring peptides whose total probability sums up to the parameter $SpectralProbability$. A spectrum is considered identified in a database if any of the generated reconstructions is present in the database. We varied the value of $SpectralProbability$, and computed the number of spectra that were identified in the *Shewanella* database and the decoy database of the same size. Table 2.1 shows the distribution of these numbers, compares them against $SpectralProbability \cdot n \cdot 50000$ (the expected number of matches in the database of size n) and shows that the number of matches in the decoy database is very close to the expected number of matches computed by MS-GF.

Figures 2.3(a,b) show the distributions of InsPecT and X!Tandem scores for the peptides identified in *Shewanella-1784* dataset against the target and decoy database. Advanced peptide identification tools are expected to have similar score distributions in target and decoy databases (otherwise, the difference between the distributions can be used to better separate the correct and false identifications). For InsPecT, the distributions in the target and decoy databases are similar, with

Kolmogorov-Smirnov (KS) distance of 0.28, indicating that InsPecT scoring cannot further differentiate between the correct and the false identifications. In case of X!Tandem E-value, there is some separation between the distributions in target and decoy database, however the distributions still have a large overlap and it is unclear what additional features can separate the correct and false identifications.

Figure 2.3(c) shows the distribution of $Energy(P, S)$ for identifications from *Shewanella-1784* dataset and demonstrates that spectral energy provides an excellent separation between the correct and false identifications. In particular, $Energy = 0$ for a significant portion of correct identifications (in these cases, the identified peptide also represents an optimal de novo reconstruction). The false identifications, on the other hand, have no identifications with $Energy = 0$. Moreover, the separation in Figure 2.3(c) indicates that the $Energy$ is complementary to many other parameters used for scoring spectra (recall that InsPecT scoring combines seven parameters but still does not attain the separation power of $Energy$). Figure 2.4 further shows the joint distribution of $SCORE$ and $Energy$ and provides an intuitive explanation why the generating function approach improves the sensitivity/specificity ratio of existing MS/MS search tools. Note that the target and decoy identifications are well separated in 2-D, with low $SCORE$ and $Energy$ for the target database and high $SCORE$ and $Energy$ for the decoy database.

Let $Score(P, S)$ be the match score of a peptide P and a spectrum S. We denote the *spectral probability* $Prob_{Score(P,S)}(S)$ of the peptide-spectrum pair (P, S) as the sum of probabilities of all peptides with match scores larger or equal to $Score(P, S)$ (when compared to S). Figure 2.3(d) shows the distribution of the spectral probability (as computed by MS-GF) for correct and false peptide identifications. This parameter also provides excellent separation between the correct and false identifications, with false identifications typically having much larger spectral probabilities $Prob_{Score(P,S)}(S)$. This is in agreement with Figure 2.3(c), further confirming that most identifications on the decoy database, in spite of their high scores, actually represent poor (sub-optimal) de novo solutions, and could be distinguished from correct solutions using MS-GF.

2.3.3 Generating functions increase the number of identifications in MS/MS database searches

Generating functions can be used to re-score the identifications obtained by various database search tools and to improve the sensitivity-specificity trade-off. We illustrate this result using *Shewanella-50000* dataset searched against the target *Shewanella* database and the decoy database using X!Tandem [49] (similar results were obtained when SEQUEST or PeptideProphet was used). The existing database search tools use two types of scores that we refer to as *raw* and *combined* scores. Raw scores (used for scanning databases) are defined by a spectrum and a peptide alone without any reference to the scores of other peptides encountered in the database search. The database-dependent combined scores integrate raw scores with other information like Δ-score of the second best peptide match (like in SEQUEST), or the distribution of scores of all peptides in the database (like in X!Tandem). We emphasize that the generating function (and the spectral probability) represents the raw score since it does not use any additional information about other peptides in the database. Below we show that the spectral probability improves on previously proposed raw scores and even outperforms the combined scores of the existing database search tools.

For each spectrum in the *Shewanella-50000* dataset, three different scores are used for analyzing the peptide identifications and constructing ROC curves: (i) X!Tandem raw score used for scanning the database, (ii) X!Tandem combined score (E-value) that integrates the raw score with the distribution of the scores for all peptides in the database, and (iii) spectral probability as reported by MS-GF for the X!Tandem identification. For each score, a varying cutoff is used, and the number of spectra that have an identification with scores above the cutoff in the *Shewanella* database and the corresponding error rate (ratio of the number of identifications on a decoy database of the same size and the number of identifications in the target database) are plotted in Figure 2.5(a).

MS-GF results in significantly higher number of identifications in the *Shewanella* database for a given error rate (number of identifications on the decoy database) when compared to the raw X!Tandem scores. Similarly, it significantly

improves on SEQUEST and PeptideProphet (data not shown). Figure 2.5(b) shows similar curves for the number of unique peptides instead of the number of spectra. For 5% error rate, X!Tandem raw/combined score identifies 1449/1613 peptides, while MS-GF identifies 1837 peptides. The advantage of MS-GF is particularly pronounced for extremely accurate identifications. For example, for 0.3% error rate (very few false identifications) MS-GF identified 1326 peptides while X!Tandem identified 943/1050 peptides with raw/combined scores. Such extremely accurate identifications are important for a notoriously difficult problem of identifying proteins based on a single peptide hit (one-hit-wonders). Indeed a single peptide hit with the error rate 0.3% may be more reliable than two peptide hits with the error rate 3% each [50, 51, 52, 46]. The fact that MS-GF has better sensitivity-specificity than even the combined X!Tandem score is surprising since MS-GF has no access to the valuable information about other peptides in the database that is incorporated into the combined X!Tandem score. We therefore argue that the spectral probability represents a valuable addition to the various "raw" scores proposed for MS/MS searches so far.

We remark that the MS-GF+X!Tandem curve in Figure 2.5 was constructed using the information about matches in the decoy database. The superior performance of MS-GF+X!Tandem over X!Tandem raises a question whether a database search based on MS-GF (i.e., using $SpectralProbability$ as a score) would be better off on its own (without using matches identified by X!Tandem). In other words, we are interested in how a database search with MS-GF scoring would fare in comparison with other database search tools. Figure 2.6 illustrates that MS-GF alone (without using X!Tandem identifications) performs better than X!Tandem. For each spectrum in the *Shewanella-50000* dataset, we generated the top-scoring peptides whose probabilities sum up to the parameter $SpectralProbability$. A spectrum is considered identified in a database if any of the generated reconstructions is present in the database. We varied the value of $SpectralProbability$, and computed the number of spectra that were identified in the *Shewanella* database and the decoy database of the same size. This essentially mimics the database search with the spectral probability as the scoring function computed by MS-GF.

Figure 2.6 provides a comparison between the number of identifications made by MS-GF and X!Tandem. Despite the fact that X!Tandem combined score utilizes information that MS-GF does not have access to, MS-GF outperforms X!Tandem. In addition, MS-GF, accurately estimates the number of hits in decoy database thus eliminating the need for the decoy database search altogether. This observation illustrates that computing scores over all possible peptides is better than observing scores over the relatively small decoy database.

Interpreting the "one-hit-wonders" is a difficult problem that often amounts to manual validations. The subjective nature of such inferences have resulted in the Proteomics Publication Guidelines to virtually discard single-hit protein identifications. In a large scale study, this inevitably results in the loss of large amounts of valuable information. For example, there are 402 proteins with single peptide hits in *Shewanella oneidensis* MR-1 [48] as opposed to 1992 proteins with multiple hits (over 20% of the expressed proteome).[6] While we estimated that nearly 75% of these "one-hit-wonders" are correct identifications (as discussed in [48, 53]), no means were available to objectively separate them from the false identifications. Below we show how MS-GF (that provides a superior separation between correct and incorrect peptide identifications for low error rates) can be used for reliable identification of the single-hit proteins.

We computed *SpectralProbability* for the peptides identified in the decoy database in [48][7]. The lowest value of $SpectralProbability$ among all these decoy identifications is 1.55×10^{-8}. Similarly, $SpectralProbability$ was computed for the peptides from the single-hit proteins and the spectral probability for 345 of them was lower than 1.55×10^{-8}. These 345 peptides represent better identifications than every identification in the decoy database, and the corresponding proteins must be considered reliably identified with virtually zero empirical error rate.[8] We further remark that many single-hit-wonders with $SpectralProbability$ below

[6]For typical bacterial MS/MS projects, the percentage of one-hit-wonders is closer to 30% (see [53]). The percentage is somewhat smaller for the unusually large *Shewanella* dataset.

[7] 1417 peptides were identified in the decoy database as compared to 28,377 peptides identified in the *Shewanella* database as described in [48]. From 1417 peptides we selected the Charge-2 and unmodified peptides for this analysis, giving 1065 peptides.

[8]See Chapter 3 for detection of sequencing errors and programmed frameshifts using a similar approach.

1.55×10^{-8} are actually more statistically significant than some proteins with multiple peptide hits but larger $SpectralProbability$ values (see [50, 51, 52, 46] for combining peptide significance scores into protein significance scores).

2.4 Discussion

While the previous approaches to evaluating the statistical significance of spectral identifications greatly improved the state of the art in peptide identification, they have not yet eliminated the decoy databases and empirical approximations from MS/MS searches. PeptideProphet [11] combines multiple scores into a single discriminant score, and fits its observed distribution to a mixture model comprising of a gaussian distribution for correct identifications and a gamma distribution for incorrect identifications. Sadygov and Yates, 2003 [12] argue that the frequencies of matches between fragment ions predicted from a random peptide and an experimental spectrum follow a hypergeometric distribution that is used to compute the probability that a peptide identification is correct. On the other hand, OMSSA tools [13] consider the same to be a Poisson distribution and accordingly compute the statistical significance of peptide identifications. These studies were taken further by Wan et al., 2006 [54] who realized the importance of generating *some* random peptides for estimating the statistical significance of the individual spectra (see also [55]) but stopped short of proposing a technique for analyzing *all* peptides. In an earlier work, Bafna and Edwards, 2003 [36] proposed an algorithm for generating suboptimal de novo reconstructions and suggested to use their score distribution for validating the optimal de novo reconstruction.

While the approaches [11, 12, 13, 22] are very valuable, neither of them rigorously solves the Spectral Matching Problem for *individual* spectra: instead they compute the error rates based on approximate fitting the empirical distributions to a standard distribution that may not carefully reflect the specifics of an individual spectrum. Moreover, they assume the same null hypothesis for *all* spectra in the sample, the assumption that may not be adequate for mass spectrometry searches. Our approach does not assume any "null hypothesis" or "noise model" for spectra

generation as in [22]. Also, it does not assume any particular approximation for the tail of the score distribution. Instead, it rigorously solves the Spectrum Matching Problem, the same problem the existing approaches attempt to solve via decoy databases and various approximations.

MS-GF allows one to accurately estimate the statistical significance of individual spectral interpretations. As described above, MS-GF can be used either to complement the decoy searches or on its own. The former case illustrates the synergy between the decoy database and the generating function approaches in cases when the generating function framework can only be applied to the results of the decoy database searches[9]. The generating function approach can be further used to generate a list of all peptides whose score exceeds a threshold and match these peptides in the protein database, thus enabling a hybrid approach to peptide identification [56, 57, 3].

While the generating function described here evaluates the statistical significance over the set of all unmodified peptides, it can be extended to analyze modified peptides in both restricted and blind [58, 59, 29] modes. The former case amounts to adding "modification edges" of fixed length while the latter case amounts to adding modification edges of arbitrary length to the amino acid graph. The dynamic programming in the resulting graph should take into account the maximum allowed number of modifications per peptide.

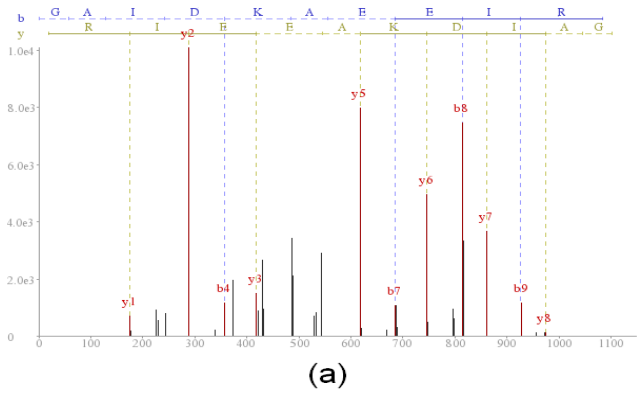
2.5 Acknowledgements

This chapter, in full, was published as "Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases". S. Kim, N. Gupta, and P. A. Pevzner. *Journal of Proteome Research*, vol. 7, no. 8, pp. 3354-3363, 2008. The dissertation author was the primary author of this paper.

[9]This is particularly relevant for estimating the error rates of *protein* identifications, re-scoring of complex non-additive scoring functions, or projects that can tolerate higher error rates (MS-GF in the database search mode becomes rather slow when high error rates are acceptable)

Table 2.1: Number of spectra in *Shewanella-50000* dataset that are identified in the *Shewanella* database (Column 2) and the decoy database (Column 3) by top peptide reconstructions with probability *SpectralProbability*. Column 4 provides the expected number of spectra that will match the decoy database given *SpectralProbability*, as computed by MS-GF without actually doing the search.

Spectral-Probability	# Correct IDs (in target DB)	# False IDs (in decoy DB)	# False IDs (predicted by MS-GF)
2e-9	8314	161	146
1e-9	7721	76	75
8e-10	7525	60	59
6e-10	7272	44	44
5e-10	7115	34	37
4e-10	6937	28	29
2e-10	6333	15	15
1e-10	5755	6	7
1e-11	3820	0	0.7



(a)

# explained b & y peaks (X)	# peptides with X explained b & y peaks (E(S,X))
13	360
12	12940
11*	146836
10	1166509
9	6946516
8	35080724
7	149410781
6	550113802
5**	1735636721

Energy=2

Delta=6

* Highest score of a database peptide
** Second highest score of a database peptide

(b)

Score (X)	# Peptides with score X	Probability of peptides with score X	Spectral Probability for score X
64	24	3.20E-11	3.20E-11
63	200	2.25E-10	2.57E-10
62	16	7.81E-14	2.57E-10
61	24	4.30E-14	2.57E-10
60*	42	1.86E-11	2.76E-10
59	510	1.36E-10	4.12E-10
58	604	7.51E-11	4.87E-10
57	316	3.04E-10	7.91E-10
56	686	4.35E-10	1.23E-09
55	956	4.30E-10	1.66E-09
54	1648	5.33E-10	2.19E-09
53	1759	9.40E-10	3.13E-09
52	3299	6.56E-10	3.79E-09
51	2996	5.15E-10	4.30E-09
50	4902	2.02E-09	6.33E-09
49	5486	1.27E-09	7.60E-09
48	7571	1.67E-09	9.27E-09
47	11772	4.64E-09	1.39E-08
46	12275	4.37E-09	1.83E-08
45	17312	5.03E-09	2.33E-08
44	18546	5.13E-09	2.84E-08
43	29686	9.03E-09	3.75E-08
42	35984	9.89E-09	4.74E-08
41	43537	9.66E-09	5.70E-08
40	51089	1.31E-08	7.01E-08
39	68199	2.16E-08	9.17E-08
38	92724	2.34E-08	1.15E-07
37	101464	2.95E-08	1.45E-07
36	197710	3.84E-08	1.83E-07
35	215618	5.23E-08	2.35E-07
34	279465	7.47E-08	3.10E-07
33	317856	8.32E-08	3.93E-07
32	398058	8.68E-08	4.80E-07
31	496905	1.32E-07	6.12E-07
30	562972	1.36E-07	7.49E-07
29	729187	1.81E-07	9.29E-07
28	852838	2.04E-07	1.13E-06
27**	1016310	2.43E-07	1.38E-06
26	1181540	3.18E-07	1.69E-06
25	1450134	2.97E-07	1.99E-06

Energy=4

Delta=33

(c)

Figure 2.1: Illustration of the generating function. (a) A spectrum S of peptide *GAIDKAEEIR* (top 43 peaks after removal of low-intensity peaks). (b) The number of peptides ($E(S, X)$) that explain X b/y peaks in this spectrum. For example, there are 360 peptides with 13 b/y ions explained ($E(S, 16) = 360$), 12940 peptides with 12 b/y ions explained, and so on. The score of the top-scoring database peptide $GAIDKAEEIR$ is 11, the optimal score among all possible peptides is 13 (such as for the peptide $QPMGAEAELR$), thus *Energy*-score is 2. The second top-scoring peptide in the database ($DQELLSEIR$) has score 5, therefore Δ-score is 6. For simplicity, a peak that explains both a b-ion and a y-ion in a particular peptide is counted as explaining two b and y peaks. (c) The (uniformly weighted) generating function of the same spectrum. The table shows the number of peptides with score X, the overall probability of peptides with score X and the total probability of all peptides with scores equal to or larger than X (spectral probability). The peptides $QIDKAEEIR$ and $QIDGAAEEIR$ represent better spectral interpretations (score 64) than the correct peptide $GAIDKAEEIR$ (score 60) resulting in *Energy*-score 4. The second best peptide in the database ($IRSIESQLR$) has score 27, therefore Δ-score is 33.

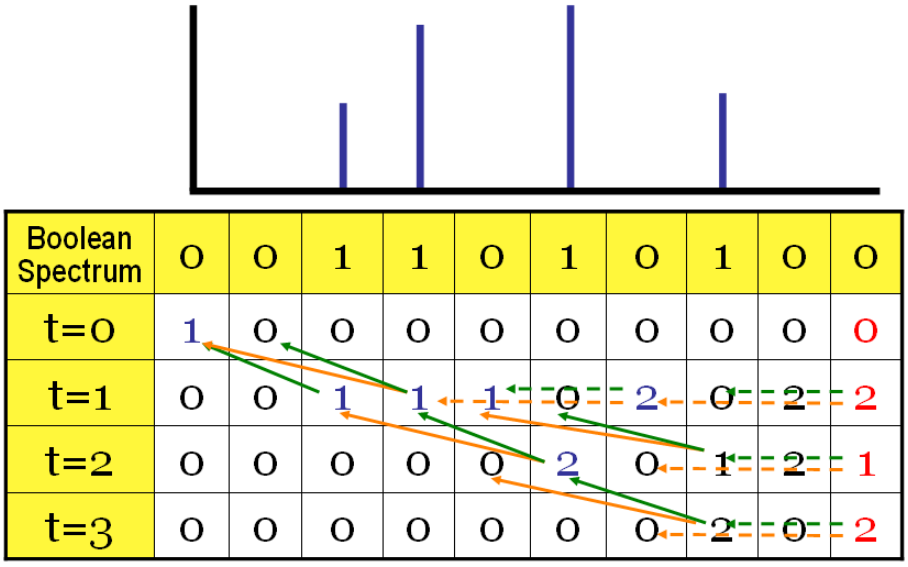


Figure 2.2: Illustration of the dynamic programming algorithm for computing the generating function. The MS-GF dynamic programming algorithm is illustrated with the help of a simplified amino acid model (only two amino acids A and B with masses 2 and 3 Daltons respectively) and a simplified discretized spectrum (only 4 peaks at 2,3, 5, and 7 Da). The scoring function used for this illustration is the number of matching prefix ions. The spectrum is converted into its boolean representation 011010100 with 1s at positions 2,3,5, and 7 (extra zero in the beginning is added to represent the variable $x(0,t)$). The vertical axis in the dynamic programming table represents scores (t). The value in each cell of the matrix represents the number of peptide reconstructions that explain the initial part of the spectrum till that position with the corresponding score. The first cell in the matrix (0,0) is initialized with 1, and the matrix is filled progressively from left to right and top to bottom. The value of each cell is computed as the sum of the values of previously filled cells which are 2 (green arrow) or 3 (orange arrows) columns before the cell under consideration. If there is a peak at the current position of the spectrum, sum is taken over the cells in the previous row, otherwise in the same row. In this example, the maximum achievable score (t) is 3, which can be obtained by two peptide reconstructions. The sequences of these reconstructions can be obtained by backtracking, as indicated by the arrows, and are found to be ABAA and BAAA. We also see that there are 2 reconstructions with score 1 and 1 reconstruction with a score of 2.

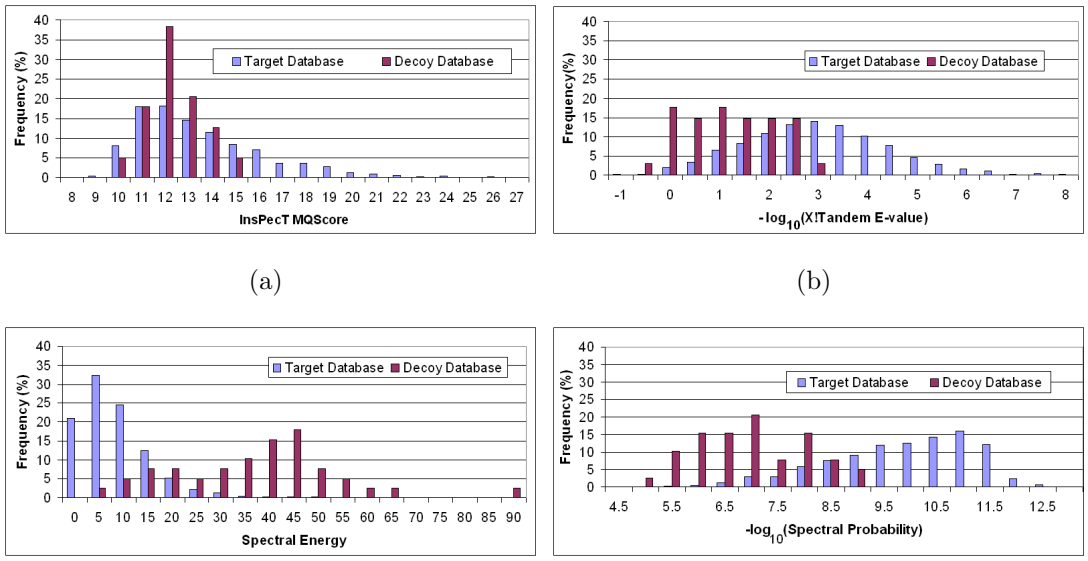


Figure 2.3: Separation between correct and incorrect identifications. Distribution of (a) InsPecT MQScore and (b) X!Tandem E-Value, for the peptides identified in *Shewanella-1784* dataset against *Shewanella* and decoy databases. X-axes show the database search scores, and Y-axes show the fraction of identifications with that score. The Kolmogorov-Smirnov (KS) distance between the two distributions is 0.28 for InsPecT scores and 0.58 for X!Tandem scores. (c) Distribution of $Energy(P, S)$ for the same dataset (the KS distance is 0.77). (d) Distribution of $-log_{10}(SpectralProbability)$ (the KS distance is 0.78). *Spectral Probability* of the pair (P, S) is defined as the sum of probabilities of all peptides whose score is larger or equal to the score $Score(P, S)$ of the match between peptide P and spectrum S.

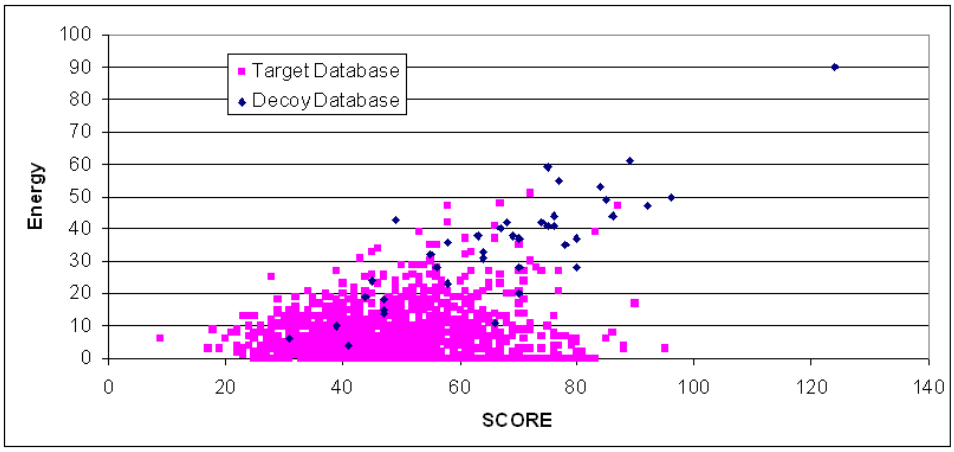
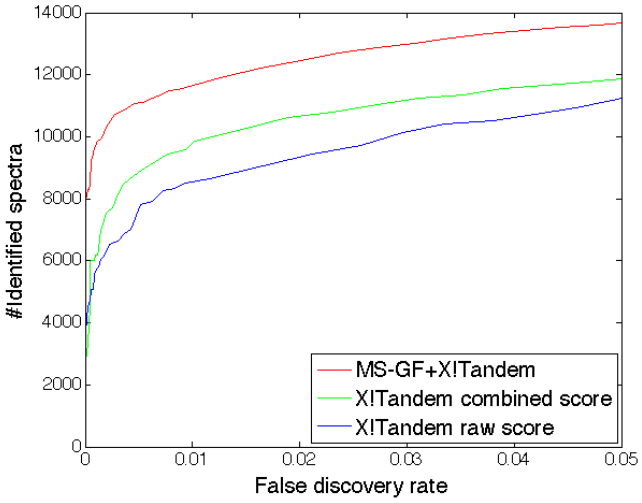
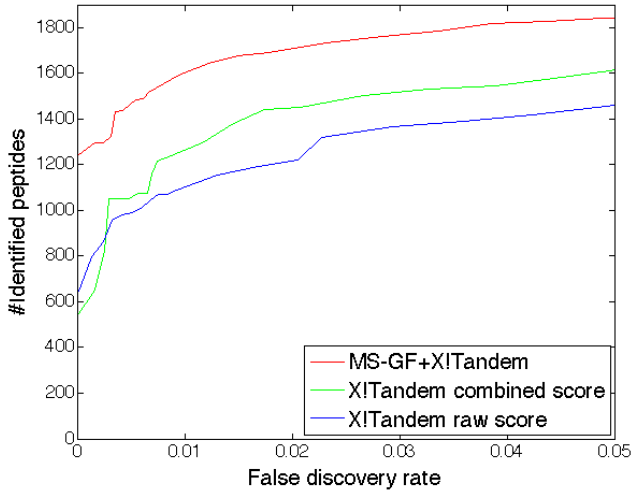


Figure 2.4: Joint distribution of *SCORE* and *Energy*. The distribution is plotted for the identifications in the *Shewanella-1784* dataset, for the peptides identified in the *Shewanella* database and the decoy database. The blue dots (decoy database) are laid over the red dots (*Shewanella* database), so that all decoy database identifications are visible.

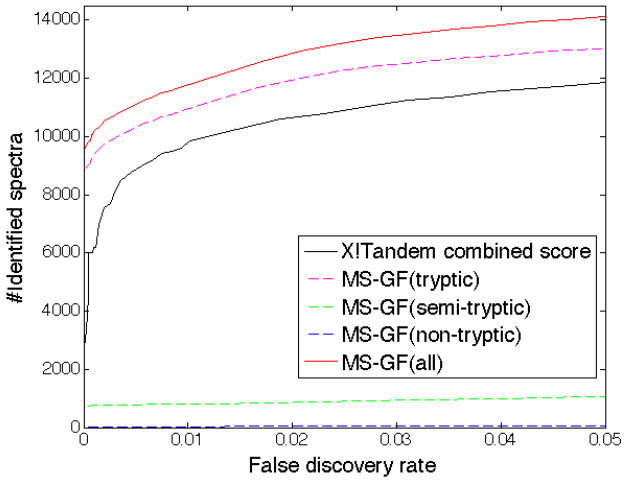


(a)

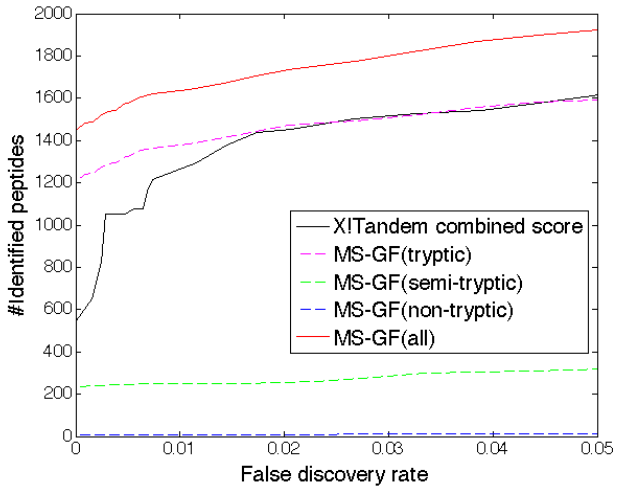


(b)

Figure 2.5: Sensitivity-specificity trade-offs. (a) Comparison of MS-GF with X!Tandem. The number of spectra identified in the *Shewanella* database and the corresponding error rate. Three scores are compared (from top to bottom): (i) *MS-GF+X!Tandem*: FPR as reported by MS-GF for the X!Tandem identifications, (ii) *X!Tandem combined score*: X!Tandem E-value that uses the raw score as well as the distribution of scores of all peptides for the given spectrum and (iii) *X!Tandem raw score*: X!Tandem hypergeometric score. (b) Similar to (a), but counting the number of unique peptides identified in the *Shewanella* and the decoy database instead of the number of identified spectra.



(a)



(b)

Figure 2.6: Performance of MS-GF vs. X!Tandem. The plots show the number of spectra identified in the *Shewanella* database and the corresponding error rate. (a) The spectral identifications in the *Shewanella* and decoy databases are divided into three groups, depending on whether the peptide endpoints are consistent with trypsin cleavage specificity: tryptic (both endpoints consistent), semi-tryptic (only one endpoint consistent) and non-tryptic (both endpoint inconsistent). The partition into these three groups illustrates MS-GF generates more tryptic peptides than the total number of peptides generated by X!Tandem.(b) Same as (a), but based on the number of unique peptides identified in each database (instead of the number of spectra). As expected, the number of peptides with both non-tryptic endpoints is very small.

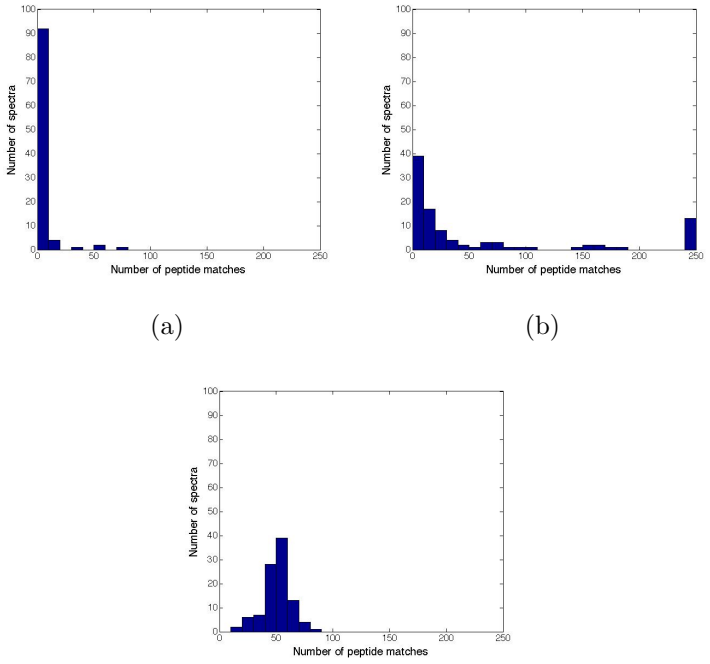


Figure 2.7: (a) Distribution of the number of peptide matches to a randomized decoy database of size 1000 times the size of the *Shewanella* database (X!Tandem search for 100 spectra with E-value 0.05). X-axis shows the number of peptide matches, and Y-axis shows the number of spectra that have these many matches. A peptide match is reported while searching the spectrum S only if it has the same or better score than the original search of S in the *Shewanella* database. (b) Similar figure for InsPecT search. Values larger than 250 on the X-axis were added to the histogram peak for value 250. (c) Similar figure for MS-GF, for which we generated the top reconstructions with the spectral probability 0.05 and counted how many of these reconstructions were found in the database. The expected number of hits in all searches is 50. InsPecT significantly overestimates the error rate, X!Tandem significantly underestimates the error rate, while MS-GF accurately computes the error rate (average number of peptide matches is 52).

Chapter 3

Integrating De Novo Sequencing with Database Search

Database search tools identify peptides by matching tandem mass spectra against a protein database. We study an alternative approach when *all* plausible de novo interpretations of a spectrum (*spectral dictionary*) are generated and then quickly matched against the database. This chapter describes a new MS-Dictionary algorithm for efficiently generating spectral dictionaries and demonstrate that MS-Dictionary can identify spectra that are missed in the database search. MS-Dictionary enables proteogenomic searches in six-frame translation of genomic sequences that may be prohibitively time-consuming for existing database searchh approaches. Such searches allow one to correct sequencing errors and find programmed frameshifts.

3.1 Introduction

In 1994, Mann and Wilm [60] proposed the *peptide sequence tag* approach and outlined its applications for protein identification. However, it took ten years for this approach to result in accurate tag-based tools like InsPecT [25] and Paragon [47], currently among the fastest MS/MS database search tools. The reason for this delay is that while generating *some* peptide sequence tags is easy, such tags are of little use unless they contain at least one correct tag with high

probability. Generating small *covering* sets of tags (i.e., the sets of tags that almost surely contain a correct tag) turned out to be a more difficult problem that was recently addressed in [25, 47, 57, 61].

Similar to generating the covering set of tags (that in most applications limited to tags of length 3), one can try to generate the covering sets of full-length peptide reconstructions that with high probability contain the correct peptide (*spectral dictionary*). Spectral dictionaries take the peptide sequence tag approach one step further by generating peptide reconstructions and ensuring that one of them is correct. They also have the potential to improve the *filtration efficiency* of tag-based tools [25, 47], for example, the filtration efficiency of 1000 de novo reconstructions of length 10 is orders of magnitude higher than even a single tag of length 3. However, while spectral dictionaries have important advantages over spectral tags, generating them remains an open problem.

The spectral dictionaries could be searched efficiently against a protein database resulting in a hybrid approach to peptide identification (Figure 3.1). While the idea of spectral dictionaries is almost as old as the idea of peptide sequence tags (Taylor and Johnson, 1997 [62]), the software tool RAId based on this approach was described only recently (Alves and Yu, 2005 [56]). However, while RAId generated promising initial results, it was based on a heuristic exhaustive search and turned out to be rather slow (2-4 minutes per spectrum) thus limiting its applicability. Also, RAId was benchmarked on a small sample thus making it difficult to evaluate its performance on large MS/MS datasets. In this chapter, we describe fast approach to generating spectral dictionaries that takes ≈ 0.1 seconds per spectrum and benchmark it on a dataset of over 20,000 peptides.

Spectral dictionaries may have an edge over the traditional MS/MS approaches in searching very large databases, e.g., six-frame translations of entire genomes. Various proteogenomic studies [63, 64, 65, 66, 67, 68, 48, 53] demonstrated that MS/MS search against a six-frame translation of the genome allows one to use MS/MS data for finding new genes, predicting programmed frameshifts, correcting DNA sequencing errors, etc. However, existing MS/MS database search tools are impractical for searches against the six-frame translation of large genomes

like human ($\approx$ 3 billion amino acids after removing repeats). Indeed, most of previous proteogenomic studies were limited to searches against the 6-frame translations of bacterial genomes. The largest proteogenomic analysis conducted so far was the search against the 6-frame translation of *Arabidopsis thaliana* that resulted in the discovery of nearly 800 new genes using InsPecT.[1] However, even fast tag-based tools like InsPecT become impractical in searches of the 20-times larger 6-frame translation of the human genome. Below we show that MS-Dictionary is able to search the 6-frame translation of human genome in roughly the same time as it takes to search the 100 times smaller database of all human proteins.

Spectral dictionaries make the size of the database almost irrelevant since the spectral dictionary can be matched against the six-frame translation as efficiently as against a much smaller database of known proteins. Since many genes remain unidentified even in the well-studied organisms (see Siepel *et al.*, 2007 [69] and Stark *et al.*, 2007 [70] for the recent discovery of over 1000 new protein-coding genes in human and fruit fly genomes), the searches in six-frame translation represent a valuable tool for proteogenomic annotations.[2]

De novo peptide sequencing represents a fast alternative to MS/MS database search. While the best de novo algorithms are orders of magnitude faster than the fastest database search tools (even on moderately sized databases), they are less accurate. However, the superior accuracy of the database search tools becomes less pronounced with the increase in the database size. Moreover, we show that for very large databases our de novo peptide sequencing algorithm compares favorably to MS/MS database search tools. Thus, searches in very large databases represent an important niche where de novo-based approaches are accurate and orders of magnitude faster than the traditional database search approaches. A number of de novo methods have been developed, including Lutefisk [62, 32], Sherenga [33], PepNovo [35], PEAKS [38], EigenMS [39], NovoHMM [40], AUDENS [41], MSNovo [43], and PILOT [42] (see also [34, 36, 72]). Most de novo

[1] Castellana, N., Payne, S., Shen, Z., Stanke, M., Briggs, S., and Bafna, V. Validation and Expansion of the Arabidopsis Gene Annotation, *submitted*.

[2] Spectral dictionaries are also helpful in searches for fusion peptides that are common in tumor proteomes but not explicitly present in protein databases (Ng and Pevzner, 2008 [71]).

tools use the *spectrum graph* approach, where a spectrum is represented as a graph with peaks as vertices that are connected by edges if their mass difference corresponds to the mass of an amino acid.

De novo peptide sequencing can also be viewed as a database search in the database of all possible peptides. Even if this time-consuming search was feasible, it would remain unclear which peptide in the database of all peptides represents the real peptide that generated the spectrum. We estimate that in 50%-95% of the cases (depending on the peptide length), the existing database search tools [25, 49, 5, 24, 73, 13] will fail to identify the correct peptide in such ultimate test since its score will be lower than the score of an incorrect peptide. We therefore argue that any de novo peptide sequencing algorithm should output multiple peptide reconstructions rather than a single reconstruction. Matching these peptides against a database results in a hybrid spectral dictionary approach that bypasses the time-consuming matching of spectra against the database.

Spectral dictionaries allow one to turn every MS/MS database search tool into a de novo peptide sequencing software (by simply running this tool on all peptides from the spectral dictionary and selecting the top-scoring peptide). After such "conversion", one can estimate how well both database search tools and de novo tools would perform on very large databases. This experiments reveals a disappointing performance of both de novo and database search tools in such an ultimate experiment: only 35%-42% of peptides of length 10 (charge 2) are correctly reconstructed in such experiments (35%, 38%, and 42% for X!Tandem, PepNovo, and InsPecT, correspondingly). Our MS-Dictionary algorithm correctly reconstructs 50% of such peptides, a significant improvement over existing approaches.[3] We further show that MS-Dictionary can search a six-frame translation of the entire human genome, the largest database ever searched for spectral interpretations.

The key problem in the spectral dictionary approach is deciding which and

[3]While MS-Dictionary compares well with X!Tandem and InsPecT for charge 2 spectra, the performance of all existing de novo tools (including MS-Dictionary and PepNovo) deteriorates for highly charged peptides (3+). The problem of de novo analysis of highly charged spectra has been recently addressed by Cao and Nesvizhskii, 2008 [74].

how many reconstructions must be generated. Generating too few peptides will lead to high false negative error rates while generating too many peptides will lead to high false positive error rates. Some de novo algorithms output a single or a fixed number (decided before the search) of peptides. For example, RAId [56] generates 1000 de novo reconstructions and matches them against a database.[4] We argue that for some spectra, generating only one reconstruction is sufficient for finding the correct peptide while in other cases (even with the same parent mass), a thousand reconstructions may be insufficient. We propose an approach for dynamically determining how many reconstructions must be generated for each spectrum, and then actually generating them.[5]

Our MS-Dictionary software (available as open source at http://proteomics.bioprojects.org/Software.html) generates spectral dictionaries based on the recently introduced concept of the *generating function* of tandem mass spectra borrowed from statistical mechanics. The generating function approach efficiently analyzes the peptide reconstructions with the optimal and sub-optimal scores and determines the statistical significance (*spectral probability* of those reconstructions (for more details, refer to Chapter 2).[6]

3.2 Methods

3.2.1 Peptide Sequencing Problem for Boolean Spectra

Dancik et al., 1999 [33] put de novo peptide sequencing in a probabilistic framework, described how to learn the parameters of the model and optimally solve it. While the Dancik model was further extended in a number of studies [35,

[4]While it may appear that matching 1000 peptides against the database is rather time consuming, the combinatorial pattern matching algorithms [75] are able to do it in negligible time.

[5] The problem of generating varying number of reconstructions for each spectrum becomes particularly important for long peptides. For instance, PepNovo [57] accurately reconstructs 54% of peptides of length 7 and only 0.4% of peptides of length 20.

[6]Although the accuracy of MS-Dictionary in the standard de novo peptide sequencing improves on the state-of-the-art tool PepNovo [35], optimizing de novo peptide sequencing is an important but not not the crucial goal for our main application. As Alves and Yu, 2005 [56] pointed out, de novo peptide sequencing and spectral dictionary approaches have similar but distinct goals: outstanding de novo algorithm is not a prerequisite for the spectral dictionary approach to perform well.

40, 76, 77], it remains unclear how to design a rigorous probabilistic model for peak intensities. We start by introducing an abstract model that seemingly has nothing to do with de novo peptide sequencing but rather describes a very general probabilistic process that transforms one Boolean string into another. We will show later that this process generalizes the probabilistic model for de novo peptide sequencing from [33] and also allows one to compute the spectral probability and the generating function of tandem mass spectra [78].

Let $s = s_1 \dots s_n$ be a Boolean string called a *spectrum* and $\pi = \pi_1 \dots \pi_n$ be a Boolean string called a *peptide*. The probability of peptide π generating spectrum s is defined as $Prob(s|\pi) = \prod_{i=1}^{n} Prob(s_i|\pi_i)$, where $Prob(x|y)$ is a 2×2 matrix (see Figure 3.2).

Given a spectrum s and a set of strings Π, we are interested in solving the problem of finding $\max_{\pi \in \Pi} Prob(s|\pi)$. Below we focus on the sets Π that are relevant to tandem mass spectrometry. Let $V = \{0, 1, \dots, n\}$ and $G(V, E)$ be a topological ordering of a DAG (Directed Acyclic Graph) such that $i < j$ for every directed edge (i, j) in E. Every path from 0 to n in G corresponds to a *G-peptide* $\pi = \pi_1 \dots \pi_n$ such that $\pi_i = 1$ iff vertex i belongs to the path (see Figure 3.2). We are interested in the following Peptide Sequencing Problem [79]:

Peptide Sequencing Problem. Given a spectrum s and a DAG G, find a G-peptide π maximizing $Prob(s|\pi)$ over all G-peptides.

In de novo peptide sequencing it is assumed that $(i, j) \in E$ iff $(j - i)$ equals the integer mass of an amino acid. Such graphs are referred to as amino acid graphs [78] (compare to *spectrum graphs* [33, 80]). As a first approximation, an MS/MS spectrum with parent mass n can be represented as a string of ones (peak present) and zeros (peak missing), with a 0/1 for every 1 Da interval. Similarly, sequences of amino acid masses (peptides) can also be represented as strings of zeros and ones. An amino acid with an integer mass α is represented as a string of $\alpha - 1$ zeros followed by a single one. Then, a peptide is simply a concatenation of Boolean strings corresponding to its amino acids. In this context, $\theta \approx 0.05$

(probability of observing a noise peak) and $\rho \approx 0.7$ (probability of observing a b-ion) represent typical values of θ and ρ for ion-trap MS/MS spectra (Figure 3.2). This somewhat simplistic Boolean model can be modified for any mass resolution, peptide fragmentation rules and peak intensities [57, 34, 36] (see below). Moreover, the more realistic model can be analyzed with exactly the same algorithm as the Boolean model [33].

The model above does not capture the fact that MS/MS spectra represent both prefix ions (b-ions series) and suffix ions (y-ions series). To reflect this we represent peptides as strings in 3-letter alphabet: 1 (theoretical b-cut), -1 (theoretical y-cut), and 0 (no cut). Given a peptide $\pi = \pi_1 \dots \pi_n$, we define its *reverse* as the peptide $\pi^* = -\pi_n \dots -\pi_1$, i.e., $\pi_i^* = -\pi_{n-i+1}$. We now redefine the probability of peptide π generating spectrum s as $Prob(s|\pi) = \prod_{i=1}^{n} Prob(s_i|\pi_i) \cdot Prob(s_i|\pi_i^*)$, where $Prob(x|y)$ is a $2 \cdot 3$ matrix

3.2.2 From Boolean spectra to MS/MS spectra

Accounting for peak intensities

While the simple model described above led to an accurate peptide sequencing algorithm [33], it does not capture the *intensities* of fragment ion in MS/MS spectra. The experimental spectra represent real-valued vectors $s_1 \dots s_n$ rather than boolean vectors (s_i is the peak intensity at mass i). One can argue that the same model based on probabilities $P(x, y)$ where x is a (real-valued) peak intensity and $y \in \{-1, 0, +1\}$ would take into account the intensities of mass-spectra. However, this model faces difficulties since (i) intensities vary between different spectra of the same peptide (ii) the value of intensity seems to be less important than the distribution of intensities over different peaks [43]. As a result, most peptide sequencing algorithms use heuristic approaches and do not try to come up with a rigorous model of spectra generation that accounts for intensities. We argue that *peak ranks* rather than peak intensities may lead to an adequate model of spectra generation. Peak ranks proved to be valuable in peptide identification, for example InsPecT [25] utilizes peak ranks in its scoring function. Below we show how to rigorously utilize peak ranks in de novo peptide sequencing and to solve

the corresponding Peptide Sequencing Problem.

We now define a spectrum $s = s_1 \dots s_n$ as a string in the alphabet $\mathcal{I}$ (ranks of peaks) and a peptide $\pi = \pi_1 \dots \pi_n$ as a string in the alphabet $\mathcal{F}$ (types of neutral losses). The probability of peptide π generating spectrum s is defined as $Prob(s|\pi) = \prod_{i=1}^{n} Prob(s_i|\pi_i) \cdot \prod_{i=1}^{n} Prob(s_i|\pi_i^*)$, where $Prob(x|y)$ is an arbitrary $|\mathcal{I}| \times |\mathcal{F}|$ matrix representing the probability that a symbol y in the peptide generates a symbol x in the spectrum.

The spectrum strings $s = s_1 \dots s_n$ are generated from tandem mass spectra as follows. For simplicity, we retain top k peaks from every MS/MS spectrum (up to $k = 150$ in our implementation). Spectra are filtered to remove noisy peaks as follows: given a peak at mass M, we retain the peak if it is among the top 5 peaks within a window of size 100 Da around M. Let's say this procedure gives t peaks, which are ranked from 1 to t. If $t > k$, we keep only the top k peaks ; if $t < k$, we re-insert the top $k - t$ peaks that were filtered out and assign them ranks $t + 1$ to k. We define s_i as the rank of the peak at mass i (if there is a peak at mass i) and define $s_i = 0$ if there is no peak at mass i.

The peptide strings $\pi = \pi_1 \dots \pi_n$ are generated from amino acid sequences as follows. We define an alphabet of fragment ions as a set of integers corresponding to neutral losses, for example ion-fragments b, $b - H_2O$, and $b - NH_3$ correspond to neutral losses $\{0, 18, 17\}$. Given a set of neutral losses $\{x_1 \dots x_t\}$, we represent every amino acid of mass α as a string $s_1 \dots s_\alpha$ of length α with $\alpha - t$ zeros and t non-zero symbols $1, 2, \dots, t$ located at positions $\alpha - x_1, \alpha - x_2, \dots, \alpha - x_t$. The peptide string $\pi = \pi_1 \dots \pi_k$ is simply a concatenation of strings corresponding to amino acids from the peptide. To make the model more accurate, we further added the doubly charged b- and y-ions as additional types of ions generated by the peptide strings.

MS-Dictionary scoring function

When applying the above model for peptide identification, we are interested in the ratio of probabilities that a spectrum is generated by a given peptide π versus probability that a spectrum is generated by a string consisting of all zeros (noise).

This can be represented as follows:

$$Prob(s|\pi)/Prob(s|0) = \prod_{i=1}^{n} Prob(s_i|\pi_i)/\prod_{i=1}^{n} Prob(s_i|0).$$

We further express it as the sum of log-odds ratios:

$$\log \frac{Prob(s|\pi)}{Prob(s|0)} = \sum_{i=1}^{n} \log \frac{Prob(s_i|\pi_i)}{Prob(s_i|0)}$$

Using the training dataset (described below), we learn $Prob(s_i|\pi_i)$ and $Prob(s_i|0)$. The learning is done separately for the lower and the higher halves of the mass range (peaks corresponding to doubly charged ions only appear in the lower part of the spectrum). A smoothing function was applied on these values for lower intensity peaks (ranks 11 to 150); for each ion type, the value at any rank was set to the average value in a window of five ranks around the given rank. These statistics vary with the length, however the differences between similar lengths (like 7 and 8) are typically small, as compared to differences between very different lengths (like 7 and 20). Thus, specific length-dependent scoring can be applied using the approximate length inferred from the parent mass of the spectrum.

The MS-Dictionary scoring function described in this chapter was compared with the scoring functions of popular database search tools, SEQUEST [24], X!Tandem [49] and InsPecT [25]. 50,000 spectra were chosen randomly from the *Shewanella* dataset and searched with Sequest, X!Tandem and InsPecT. The score of the best peptide for each spectrum from database search was compared with MS-Dictionary score for the same spectrum-peptide pair. We find good correlation between the MS-Dictionary scoring function and the scoring functions used in the database search tools, the correlation coefficients being 0.87 for SEQUEST, 0.90 for X!Tandem, and 0.96 for InsPecT (Figure 3.3). These correlations are even better than the correlation between the database search tools themselves (for example, InsPecT and X!Tandem raw scores have a correlation coefficient of only 0.75).

Suboptimal peptide reconstructions

We use the dynamic programming algorithm for computing the spectral probability and the generating function described in Chapter 2. The number of

peptide reconstructions is computed for each mass value, and the optimal score is determined for a mass within specified error tolerance from the $PrecursorMass$. We then generate top reconstructions such that their $SpectralProbability$ (see Chapter 2 for details) adds up to a fixed threshold (we typically use 10^{-9}). Starting from the topmost score, reconstructions at each score are selected until their cumulative probability exceeds the threshold (all reconstructions at the borderline score are selected; hence the total probability may marginally exceed the threshold). We limit the number of reconstructions generated for any spectrum to at most 100,000.

The dynamic programming table is constructed for all mass values between 0 and $PrecursorMass + 0.5$, with a resolution of 0.1 Da. The number of reconstructions is computed by summing up the results for all mass values in a window of 1 Da around the exact $PrecursorMass$, to account for the low-accuracy of ion-trap mass spectrometers. In case of precision mass-spectrometry (e.g. FTMS), accurate solutions (with low parent mass error) can be obtained by increasing the resolution and reducing the size of the window around the $PrecursorMass$. For efficient computation, I and L are treated as the same amino acid, resulting in 19-letter amino acid alphabet at the time of generating reconstructions. In the low accuracy setting, Q and K are also treated as the same amino acid.

Symmetric versus anti-symmetric de novo reconstructions

Some de novo reconstructions may be *symmetric*, i.e., the same peak in the spectrum may contribute to the score up to four times, as a singly charged or doubly charged, b-ion or y-ion. The algorithm to alleviate this problem was proposed by Chen et al. [34] and further improved in [38, 36]. Later Lu and Chen, 2003 [81] designed an algorithm for generating all anti-symmetric peptide reconstructions. We have chosen not to use the anti-symmetric path approach in MS-Dictionary since (i) it leads to a significant time overhead when many reconstructions are generated and (ii) it does not take into account doubly-charged ion fragments that often have high intensities and thus contribute significantly to MS-Dictionary scores. To accurately score the symmetric reconstruction, MS-Dictionary re-scores

the obtained peptide reconstructions to exclude multiple contributions from the same peak. Starting with the highest scoring reconstructions, we check the peptide sequence to determine if there are any peaks that have multiple contributions to the score. These peptides are re-scored by using only the largest contributions from such peaks.

Template-Free Spectral Recalibration

Recalibration of tandem mass spectra is important for correcting systematic mass errors. All existing spectral recalibration tools use *templates* (interpreted spectra with known b/y peaks) to perform linear recalibration using either least squares fit [32, 38, 82] or least median of squares fit [39]. In the de novo peptide sequencing framework the reliable templates are hard to obtain thus reducing the utility of spectral recalibration to QTOF and LTQ-FT data. In the low mass-accuracy setting, the applications of template-based spectral recalibration are mainly limited to validating candidate peptide identifications. As a result, de novo peptide sequencing programs commonly default to a rather high fragment mass tolerance (e.g., 0.5 Da for ion-trap data) and thus result in many erroneous spectral interpretations. We describe a *template-free* spectral recalibration procedure for ion-trap mass spectra and demonstrate that it reduces the required mass tolerance from 0.5 Da to 0.2 Da. We further show that this recalibration leads to significant improvement in MS-Dictionary accuracy.

The fractional masses of amino acids may be as large as 0.1 for Arginine (mass 156.1 Da). The first step of our MS-Recalibration tool is *rescaling* all peaks in the spectrum by multiplying all masses by 0.9995 to minimize the theoretical fractional masses of amino acids. After rescaling the fractional mass of Arginine is 0.02 (156.02 Da) and the fractional masses of all other amino acids are below 0.04 (the average fractional mass reduces three-fold from 0.06 to to 0.02).

MS-Calibration further filters the rescaled spectra to retain the high intensity peaks using a sliding window as described above. Using $Int(m)$ and $Frac(m)$ to denote the integer and fractional part of mass m (respectively), our goal is to find α and β minimizing the sum $\sum(Frac(\alpha \cdot m + \beta))^2$ over all masses m in the

rescaled filtered spectrum (Figure 3.4(a)). The coefficients α and β are computed with the least squares fit algorithm and are used to recalibrate all peaks in the rescaled spectrum. While MS-Recalibration has no information about the peptide that produced the spectrum, Figure 3.4(b) illustrates that it achieves almost the same accuracy as the template-based approaches that recalibrate the spectra based on the information about the correct positions of b/y ions. After applying MS-Recalibration, one can safely set the mass tolerance to 0.2 Da (and retain 96% of b/y peaks) as compared to the 0.5 Da in existing approaches. Another advantage of our method is that it makes the mass error distributions centered around zero regardless of their positions in the spectrum. This feature is important for designing a new scoring function that carefully account for errors in peak positions (see below).

Incorporating Mass Errors Into the Scoring Function

Most de novo peptide sequencing tools [57, 62, 32, 33, 38, 39, 40, 43, 42, 36, 83, 35, 84] setup a fixed mass error threshold (e.g., 0.5 Da for ion-traps) and compute the scoring functions for all peaks within this error threshold. Bafna and Edwards, 2001 [85] and Mo et al., 2007 [43] noticed that assigning the same scores to *all* peaks within the error threshold may not be the optimal way to score spectra in both database search and de novo peptide sequencing applications. For example, a high intensity peak with mass error 0.5 Da is typically less "reliable" than a medium intensity peak with mass error 0.1. Recent incorporation of mass errors into the scoring function (as a quantitative component rather than a cut-off) led to a significant improvement in MSNovo accuracy [43]. MS-Dictionary also incorporates mass errors in the scoring functions and further improves MSNovo model as described below.

MSNovo used unified peak error model (Gaussian distribution) and peak rank model (exponential distribution) independent on the ion type, rank and position of each peak. However, Figure 3.5(a) illustrates that different fragment ions have different error models. Figure 3.5(b) reveals that peak ranks and mass errors (that are assumed to be independent in MSNovo) are strongly correlated. Also,

Figure 3.5(b) reveals subtle irregularity in noise peaks indicating that the noise model in Mo et al., 2007 [43] needs to be adjusted. MS-Dictionary takes these observations into account and incorporates the mass errors into its scoring function using a more adequate error model than Mo et. al.,2007 [43]. Below we briefly describe the error-dependent scoring for boolean spectra (this model can be extended to MS/MS spectra as described above).

The boolean spectra model assumes that a peptide symbol π_i generates the spectrum symbol s_i at exactly the same position. We now extend this model by assuming that the peptide symbol π_i can generate spectrum symbol $s_{i+\epsilon}$, where ϵ represents a *mass measurement error*. We assume that errors are "small", i.e., they do not exceed a threshold ϵ_{max} (ϵ_{max} is typically 0.5 for ion-trap spectra). Incorporating errors into the spectrum generation model requires introducing the 3-dimensional matrix $Prob(x, \epsilon|y)$, where $-\epsilon_{max} \leq \epsilon \leq +\epsilon_{max}$ and x and y are boolean as before. The probability of peptide π generating a spectrum s with error $\epsilon = \epsilon_1, \ldots, \epsilon_n$ can now be defined as $Prob(s, \epsilon|\pi) = \prod_{i=1}^{n} Prob(s_{i+\epsilon_i}|\pi_i) = \prod_{i=1}^{n} Prob(s_{i+\epsilon_i}, \epsilon_i|\pi_i)$. The Peptide Sequencing Problem can now be reformulated as follows:

Peptide Sequencing Problem with Errors. Given a spectrum s and a DAG G, find a G-peptide π and mass errors ϵ maximizing $Prob(s, \epsilon|\pi)$ over all G-peptides and over all mass errors ϵ.

The matrix $Prob(x, \epsilon|y)$ was learned from the training sample and the learned parameters were further used in the dynamic programming algorithm as described before. Table 3.1 compares the performance of MS-Dictionary with PepNovo version 1.03 and illustrates that MS-Dictionary outperforms PepNovo for all peptide lengths.

3.3 Results

3.3.1 Datasets

We used the previously published *Shewanella oneidensis MR-1* spectral dataset containing 14.5 million spectra. The experimental procedures[7] for acquiring the spectra and identifications from this dataset are described in [48]. 28,377 peptides were reliably identified with false discovery rate 5% using InsPecT (spectrum-level FDR is 1%). InsPecT search was run using default parameter settings (fragment ion tolerance of 0.5 Da and parent mass tolerance of 2.5 Da). For this study, we selected 21,087 tryptic peptides with charge 2, obtained one representative spectra for each of these peptides (most peptides were identified from multiple spectra), and grouped these by the length of their peptide identifications to form a test dataset for each length. We will refer to the length of the InsPecT identification of a spectrum as the *spectrum length*. For the sake of convenience, all lengths 7 through 10 and even lengths between 10 and 20 were considered. The trends across these lengths show smooth progression and there is no reason to believe that the odd lengths between 10 and 20 would show any deviant behavior. To avoid computational artifacts introduced by errors in the parent mass, we have chosen to correct the parent masses according to the InsPecT identifications.

3.3.2 Generating multiple de novo reconstructions

A spectrum may have many reconstructions with the optimal score, and in these cases, reporting only one reconstruction is clearly deficient. For example, Figure 3.6 shows a spectrum for which two distinct peptides, LHEALPDPEK and HLEALGAFYK, receive the optimal de novo score of 90.

We further argue that even generating all optimal reconstructions may not be sufficient for finding the correct peptide. For many spectra, the correct peptide has a lower score than an incorrect peptide. Figure 3.7 shows a spectrum for

[7]The spectra were acquired on an ion trap MS (LCQ, ThermoFinnigan, San Jose, CA) using electrospray ionization (ESI). The program extract_msn (ThermoFinnigan) was used to generate the dta files with standard options.

which the correct peptide FINVIMQDGK (as identified reliably by InsPecT) has score of 111, a high score that exceeds the average score of correct identifications. However, another reconstruction YPNVMLQDGK (not present in the database) has an even higher score 123. We note that for $\approx 60\%$ of length 10 spectra, the correct peptide has suboptimal PepNovo score ($\approx 50\%$ for MS-Dictionary score), and this fraction quickly increases with the peptide length (Figure 3.8). Since the existing de novo approaches fail to identify the correct peptide as the optimal reconstruction in a large fraction of the spectra, a de novo method should consider multiple reconstructions with sub-optimal scores.

3.3.3 How existing database search approaches fare while searching very large databases?

All database search tools we tested would fail to identify the correct peptide for more than half of the length 10 spectra if they were searching through the database of all possible peptides. This is an indication of limitations of the scoring functions of existing database search tools. Since actually searching a database of all peptides is impractical, we conservatively estimate the error rates of MS/MS database search tools by constructing a custom database for each spectrum containing all de novo reconstructions with MS-Dictionary scores better or equal to the correct peptide. Even if we used the theoretical database of all possible peptides, it is likely that the identified peptides would be one of those top reconstructions that we included in our custom database. The rate of finding the correct peptide would only drop if more peptides were added. InsPecT was able to identify the correct peptide (peptide identified in the the *Shewanella* database in [48]) on such custom database in only 42% of the cases and X!Tandem in 35% cases for length 10 peptides. Both InsPecT (version 2006.09.07) and X!Tandem (version 2007.01.01.2) were run with parent mass tolerance of 2.5 Da, fragment mass tolerance of 0.5 Da, fixed modification of C+57, no optional modifications and without any enzyme preference. The best match for each spectrum is reported. The parent masses of spectra were corrected according to the mass of the correct peptide. Table 3.2 illustrates that the accuracy of various tools decreases sharply with the increase

in the spectrum length. PepNovo (a de novo search method) has similar or better accuracy than InsPecT in finding the correct peptide reconstruction. PepNovo version 1.03 was used with fixed C+57 modification.

We remark that in some applications (e.g., search in large EST databases or using MS/MS for proteogenomic annotations [68, 48]), the databases are very large. It implies that search in such databases (at least for shorter peptides) is not unlike the search in the database of all peptides. Table 3.2 leads to a surprising conclusion that for short peptides simply generating de novo reconstructions and matching them against the database may be more accurate (and much faster) approach than X!Tandem/InsPecT in case of very large databases. Below we show that MS-Dictionary leads to a better performance than InsPecT, X!Tandem, and PepNovo in such applications (Figure 3.9).

3.3.4 Performance of MS-Dictionary

The test datasets (all peptide identifications in *Shewanella*) were analyzed with MS-Dictionary for each peptide length. The size of the spectral dictionary depends on the $SpectralProbability$ parameter of the generating function [78] that influences the error rate of peptide identifications if the spectrum was submitted to a database search. Since we deal with tryptic peptides, we only consider the reconstructions that end in K or R (although MS-Dictionary is not limited to tryptic peptides).[8]

As the spectrum length increases, the size of the peptide search space increases dramatically, making it harder to generate the spectral dictionary. Thus all de novo search methods yield lower accuracy for longer peptides. The generating function approach allows one to dynamically determine the number of peptide reconstructions and increase the chance of finding the correct peptide in the set of de novo reconstructions (see Figure 3.9).

The number of reconstructions obtained for these same-length spectra varies over orders of magnitude. While the peak of the distribution of the number of

[8]While this analysis looses peptides at the C-terminus of proteins, it will have a minor effect on the reported statistics.

reconstruction is at $log_2(size\ of\ spectral\ dictionary) \approx 10$ (comparable to the number of reconstructions generated in [56]), some of these spectra have fewer than 100 or more than 10,000 reconstructions. This remarkable variance in the size of spectral dictionaries illustrates the point that different spectra have different number of plausible reconstructions and raises a concern about de novo methods that return a fixed number of peptides.

Recently, Frank *et al.*, 2007 [86] described de novo peptide sequencing for data acquired from FT-ICR instruments when *both* the parent mass and the peak positions are accurate. However, acquiring such spectra remains time-consuming, and an intermediate approach that is gaining prominence is to acquire mass spectra with high precision at MS1 stage and lower precision at MS/MS stage, giving accurate parent mass but inaccurate peak positions. However, the existing de novo search methods are aimed toward ion traps or other low accuracy mass spectrometers, that may have parent mass errors on the order of 1 Dalton. Since vertices in the spectrum graph are constructed based on low accuracy peaks, it is not clear how to exploit the accurate parent mass information that is available from new high accuracy instruments. Availability of accurate $PrecursorMass$ values can be effectively utilized in MS-Dictionary to filter the reconstructions. The number of reconstructions for 5 ppm accuracy is typically 4-16 times smaller than the corresponding numbers for 0.5 Dalton accuracy (data are not shown).

3.3.5 Using MS-Dictionary for database search

In any database search, a large number of spectra remain unidentified. This may happen due to several reasons: these spectra may have many missing or noisy peaks making them difficult to interpret, the corresponding peptide may not be present in the database or the peptide may have a post-translational modification not captured by the search algorithm. In case of *Shewanella oneidensis* MR-1, only $\approx$ 10% of the 14.5 million spectra were reliably identified [48]. We show that MS-Dictionary is able to find identifications for some previously unidentified spectra.

We selected all ($\approx$ 600 thousands) spectra of charge 2 from the *Shewanella*

dataset within the $PrecursorMass$ range from 1100 to 1200 Da (the typical mass range for length 10 peptides). All these spectra were searched against *Shewanella* proteome with MS-Dictionary (generated with spectral probability 1e-9), InsPecT, and X!Tandem. The same analysis was repeated with a decoy database of the same size. A spectrum is considered identified if any of the reconstructions is present in the six frame translation of the *Shewanella* genome (target database). Figure 3.10 demonstrates that InsPecT and MS-Dictionary significantly improve on X!Tandem (at 5% FDR, X!Tandem, InsPecT and MS-Dictionary identified 3272, 4184 and 4137 peptides respectively). We further rescored InsPecT identifications using MS-GF spectral probabilities achieving an even better performance for the hybrid InsPecT $\oplus$ MS-GF hybrid tool (4299 peptide identifications at 5% FDR). Figure 3.11 shows the Venn diagrams of peptides identified by X!Tandem, InsPecT, MS-Dictionary, and InsPecT $\oplus$ MS-GF.

To further illustrate applicability of MS-Dictionary in proteogenomics applications we extended the analysis of *Shewanella* proteome described above (Figure 3.10) to the seven times larger 6-frame translation of *Shewanella*. We selected all spectra from *Shewanella* dataset, that were not identified in the InsPecT database search, with the $ParentMass$ range from 1100 to 1200 Da and with MS-GF scores above 50 (24,814 spectra).[9] MS-Dictionary generated spectral dictionaries for these spectra, at three different values of $SpectralProbability$. The same analysis was repeated with a decoy database of the same size. A spectrum is considered identified if any of the reconstructions is present in the six frame translation of the *Shewanella* genome (target database). Table 3.3 shows the number of new peptides identified by MS-Dictionary in each database that were not found in the earlier database search. For $SpectralProbability = 10^{-10}$, 1007 new peptides are identified from 6211 spectra in the target database, while only 6 peptides (from 6 spectra) are identified in the decoy database, corresponding to a peptide level false discovery rate of 0.6%. As the $SpectralProbability$ is lowered, the false

[9]While most spectra with MS-GF scores above 50 correspond to high quality peptide identifications by both InsPecT and X!Tandem, a significant portion of them may have borderline InsPecT/X!Tandem scores. As discussed in Chapter 2, such low scores may reflect deficiencies of the underlined scoring approaches.

discovery rate turns into zero at 2×10^{-11} with 794 peptide identifications. 280 of them were previously identified by InsPecT (from other higher-quality spectra) but 514 represent new peptide identifications. Interestingly, 512 (99.6 percent) of them map to the known protein sequences (including contaminants), providing further confirmation that these identifications are correct. Indeed, since the size of the *Shewanella* protein database is only $\approx$ 15% of the size of six-frame *Shewanella* translation, one expects that only 15% of these proteins would hit the *Shewanella* database by chance. Moreover, out of 512 peptides, 508 are matched to expressed proteins (confirmed by at least two InsPecT identifications in [48]) and 2 are matched to proteins with a single identified peptide, confirming the expression of these proteins.

A closer look at the two peptides that fall outside the annotated proteins reveals two frameshifts. The first peptide, IAVGLSSANFGR, maps downstream of the gene SO_2754 which is annotated as "hypothetical sodium-type flagellar protein MotY", and has length 122 aa. BLAST [14] query of the peptide against other *Shewanella* strains shows that the peptide is conserved in four other strains and contained in longer proteins of length 289. By aligning the nucleotide sequence of *Shewanella oneidensis* MR-1 against these other strains, we find a sequencing error (insertion of an extra A at nucleotide position 362) that results in a stop codon and early truncation of the gene with only 122 amino acids. The second peptide SDIGWGSQIR falls in the region of the gene SO_0991 (peptide chain release factor 2) which is now annotated in TIGR as a programmed frameshift (but has the correct protein sequence missing from fasta files because of the frameshift). These examples show that new peptide identifications from MS-Dictionary not only increase coverage for annotated genes but also provide clues for correcting gene annotations.

We note that peptide identifications reported here based on the spectra in 1100 to 1200 Da $PrecursorMass$ range only, and their number is expected to be much larger if spectra of other masses are also included. Spectra in lower or higher mass ranges also show similar trends as spectra in the 1100 to 1200 Da range (data now shown). MS-Dictionary, thus, has the potential to provide a

significant number of new peptide identifications from spectra that were missed in the traditional database searches.

3.3.6 Searching the six-frame translation of human genome with MS-Dictionary

While mass spectrometry have been successfully used for bacterial gene predictions [63, 64, 65, 66, 48, 53, 87], the proteogenomic studies of large eukaryotic genomes are still in infancy. Even the fastest MS/MS database search tools become impractical in such studies since they require searches in huge databases resulting from the 6-frame translations of eukaryotic genomes ($\approx$ 2.5 billion amino acids for repeat-masked human genome). Tanner et al., 2007 [68] and Edwards, 2007 [88] made a step towards proteogenomic searches of human genome by combining the EST and MS/MS analysis. While this approach is very valuable it can only be successful if the same exons are supported by both EST and MS/MS data. The largest proteogenomic analysis conducted so far is the search of the six-frame translation of *Arabidopsis thaliana* that resulted in the discovery of nearly 800 new genes using InsPecT.[1] While InsPecT is 10 times faster than X!Tandem and 60 times faster than SEQUEST (Payne et al., 2008 [89]), it becomes too slow in searches of the translated mammalian genomes. Since neither InsPecT, nor X!Tandem can search the translated human genome,[10] we ran InsPecT on a 124 times smaller database and assumed that its running time is proportional to the database size. The running time of InsPecT is estimated at 42 seconds per spectrum[11], while MS-Dictionary takes less than 1 second per spectrum on average on a desktop machine with 2.16Ghz Intel processor. Below we demonstrate that MS-Dictionary can search the translated human genome and identify over 10,000 human peptides with low FPR. Recently, Tanner et al., 2007 [68] demonstrated that such peptides can significantly improve the accuracy of traditional *de novo* gene prediction tools and boosted the accuracy of GeneID predictions by 0.65 correct exons per gene on

[10]Both tools report unexpected errors on the translated human genome.

[11]This is a lower bound that does not account for overhead caused by indexing/partitioning of large databases.

average.

MS-Dictionary generates the spectral dictionary for each spectrum and uses fast pattern matching to match the spectral dictionary against the indexed database.[12] We used a simple partitioning/indexing that divides the translated human genome into 124 equally sized sub-genomes. Generating a spectral dictionary with 10,000 reconstructions takes 0.1 seconds per spectrum and pattern matching of a spectral dictionary against all 124 databases (including I/O overhead) takes 0.8 seconds per spectral dictionary on average. This results in less than 1 second running time, a 40-fold speed-up over InsPecT.[13]

To benchmark MS-Dictionary we used the human HEK293 MS/MS dataset generated in Steve Briggs' lab. We focus on 48,926 doubly-charged peptides with tryptic C-terminus identified by InsPecT[14] (InsPecT version 20070613, human IPI database version 3.18) with 2.5% false discovery rate (for detailed description see [68, 90]). We removed 17,821 peptides that span the exon boundaries (these peptides cannot be identified by searching the translated human genome) resulting in 31,105 peptides. Since most peptides in HEK293 are represented by multiple spectra, we randomly selected one spectrum out of all spectra of the same peptide. We further searched 31,105 spectra against the translated human genome (version 48 from Ensembl ftp server ftp://ftp.ensembl.org) with masked repeats and with corrected parent mass as described before. For each spectrum, we generated a spectral dictionary with $SpectralProbability = 10^{-11}$ and limited the maximum size of spectral dictionaries to 10,000. Each peptide in the spectral dictionary was matched (without errors) against the translated human genome.

The searches in the translated human genome are not expected to identify all spectra reliably identified in the human protein database. Indeed, Castellana et al., 2008[1] "lost" $\approx$ 30% of all identifications of peptides falling within exons after switching from protein database to the translated genome database of *Arabidopsis thaliana*. Such losses are unavoidable since many reliable identifications in

[12]Indexing the entire 6-frame translation of the human genome takes less than an hour.

[13]We estimate that optimized indexing/partitioning or running MS-Dictionary on a large shared memory machine would further reduce the running time.

[14]While MS-Dictionary generates both tryptic and non-tryptic peptides, we selected doubly-charged peptides with tryptic C-terminus to simplify benchmarking.

the protein database turn into statistically insignificant identifications in the much larger translated genome. For example, while the $SpectralProbability = 10^{-10}$ makes sense for searching the human protein database, it results in very high error rates (FPR=25%) in a ≈ 100 times larger translated human genome. Therefore, all peptide identifications with $SpectralProbability \geq 10^{-10}$ will be lost after switching from the protein database to the translated human genome.[15] We have therefore chosen $SpectralProbability = 10^{-11}$ as a threshold resulting in estimated FPR=$DatabaseSize \cdot SpectralProbability = 2.5 \cdot 10^9 \cdot 10^{-11} = 0.025$. Since 9,470 out of 31,105 peptides (30%) have $SpectralProbability$ exceeding 10^{-11}, they cannot be identified in any sensible database search against the translated human genome. It leaves us with 21,635 peptides that can be potentially identified in the translated human genome.

MS-Dictionary identified 10,266 out of 21,635 spectra in the translated human genome. 98.9% of the identified peptides fall into the human proteins and only 1.1% fall into non-coding regions.[16] To further estimate FPR of our experiment, we selected a single run (25,746 spectra), picked out unidentified doubly-charged spectra in this run (16,205 spectra), and used MS-Dictionary to generate spectral dictionaries and match them against the translated human genome. MS-Dictionary identified only 71 spectra in this experiment, corresponding to FPR 0.44%.

Therefore, MS-Dictionary reliably identifies $\approx 10,000$ peptides from human proteins *without* knowing the human proteome. However, it also "looses" $\approx 11,000$ peptides that can be potentially identified in searches of the translated human genome. Figure 3.12 illustrates that while MS-Dictionary identifies a large fraction of peptides of length 10-13, the performance deteriorates for shorter and longer peptides. Since the $SpectralProbability$ threshold has to be low in proteogenomic applications, only very high quality spectra of shorter peptides represent reliable identifications (only 23% of spectra of length 9). This does not indicate the poor performance of MS-Dictionary but rather reflects the stringent

[15] In particular, all peptides of length 8 and shorter are likely to be lost since $SpectralProbability$ even of a single peptide of length 8 is rather high ($\approx 0.4 \cdot 10^{-10}$).

[16] While most spectral dictionaries have zero or one hit in the human genome, 1.8% of them have multiple hits (typically 2 hits).

threshold. For the spectra of length more than 14 aa, the performance of MS-Dictionary deteriorates because of the limited size of spectral dictionaries. Further algorithmic developments (e.g. generating dictionaries of long tags) are needed to address this shortcoming of MS-Dictionary.

3.4 Discussion

In this chapter, we demonstrate the importance of obtaining multiple de novo peptide reconstructions and describe MS-Dictionary tool for generating these reconstructions. We emphasize that the number of generated reconstructions must not be fixed *a priori*, as done by existing de novo tools, but decided dynamically for the given spectrum since the number of plausible reconstructions varies from spectrum to spectrum. We use the generating function approach [78] that allows one to determine the set of reconstructions that must be reported. The ability to generate spectral dictionaries makes this method useful for hybrid de novo based database search, by increasing the likelihood of finding the correct peptide while keeping the number of false identifications low. MS-Dictionary identifies new peptides from spectra that were not identified with regular database search. MS-Dictionary can be modified to search for mutations and polymorphisms by simply substituting the exact pattern matching by error-tolerant pattern matching of spectral dictionaries against databases.

Future work will focus on developing this hybrid approach into a viable tool for peptide identification by extending it to highly-charged spectra and improving the efficiency of this approach in case of longer peptides. Deteriorated performance for highly-charged and long peptides is an important limitation of all de novo approaches to spectral interpretations. The existing de novo peptide sequencing tools are aimed at charge 2 peptides with the single exception of GBST algorithm [72] that is best suited for tag generation rather than full length de novo peptide sequencing, the focus of this chapter. All tools we tested also deteriorated while searching longer peptides in very large databases. For example, InsPecT and X!Tandem would correctly identify only 16% and 11% of all length 14 peptides

in the de novo peptide sequencing framework (Table 3.2). While MS-Dictionary improves on these tools, its accuracy is also rather low (18%). This observation reveals the shortcomings of existing de novo and database search tools that often score the incorrect peptides higher than the correct peptides. Frank et al., 2007 [86] recently discussed the "homeometric peptides" that represent the key obstacle for developing better de novo algorithms (they become more pronounced with the increase in the peptide length). This problem is partially alleviated by generating all reconstructions with a given $SpectralProbability$ and further matching them against a database (Figure 3.9).

3.5 Acknowledgements

This chapter, in full, was published as "Spectral Dictionaries: Integrating De Novo Peptide Sequencing with Database Search of Tandem Mass Spectra". S. Kim, N. Gupta, N. Bandeira, and P. A. Pevzner. *Molecular & Cellular Proteomics*, vol. 8, no. 1, pp. 53-69, 2009. The dissertation author was the primary author of this paper.

Table 3.1: Comparison of MS-Dictionary and PepNovo reveals that MS-Dictionary outperforms PepNovo for all peptide length (Shewanella dataset).

	% correct amino acids		% correct peptides	
Length	PepNovo	MS-Dictionary	PepNovo	MS-Dictionary
8	88.7	92.2	51.1	58.1
10	85.8	91.2	38.2	49.6
12	79.7	87.2	23.1	34.5
14	71.1	81.7	11.8	17.8
16	61.1	79.0	3.8	12.9
18	56.8	74.2	1.5	7.6
20	49.8	65.6	0.3	3.3

Table 3.2: Accuracy of InsPecT and X!Tandem against a database of all peptides, estimated as the percentage of spectra for which the correct peptide will be identified with maximal score in the database search. PepNovo and MS-Dictionary accuracy (percentage of spectra for which the correct peptide is a top-scoring peptide) is added for comparison. Peptides that differ by amino acid substitutions I/L and K/Q with similar masses are considered valid reconstructions.

Length	InsPecT	X!Tandem	PepNovo	MS-Dictionary
7	63	51	54	57
8	59	47	51	58
9	48	41	45	51
10	42	35	38	50
12	18	22	23	35
14	16	11	12	18

Table 3.3: MS-Dictionary identification of *Shewanella* spectra that were not identified in the InsPecT search in [48]. For different values of $SpectralProbability$ (first column), the number of peptide identifications (IDs) on the target database (second column) and the decoy database (third column) are reported. The numbers in parentheses represent the corresponding number of spectral identifications (many spectra correspond to the same peptide identification). The target database here is the six-frame translation of the whole *Shewanella* genome containing ≈ 10 million aa, and a decoy database of the same size is used. The fourth column provides the false discovery rate (FDR) at the peptide level (ratio of decoy and target database peptide IDs), and the fifth specifies the number of new peptides identified in the target database that were not observed in the InsPecT search. The number in parenthesizes in the last column shows the number of new peptides mapped to the protein-coding regions and illustrates that while the protein database is only 15% of the size of the six-frame translation, 97.1%-99.6% of these peptides are mapped to the protein database. These peptides are missed by InsPecT either due to borderline p-values (as shown in [78], the generating function of MS-Dictionary results in better separation between correct and erroneous hits than the scoring functions of InsPect and X!Tandem) or due to absence of good peptide sequence tags.

$SpectralProbability$	IDs(target)	IDs(decoy)	FDR	New Peptides
1e-9	1169(8771)	29(64)	0.025	768(746)
1e-10	995(6171)	6(6)	0.006	652(646)
5e-11	914(5327)	2(2)	0.002	595(591)
2e-11	794(4269)	0(0)	0	514(512)

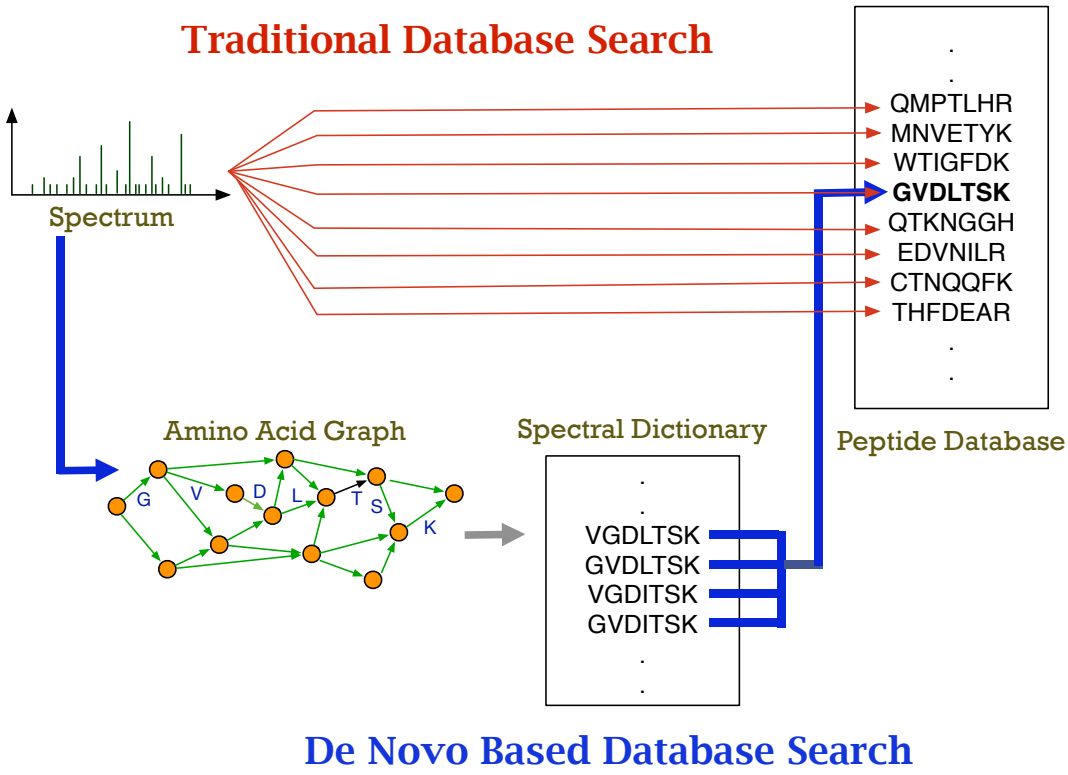


Figure 3.1: Two approaches to peptide identification: traditional approach based on comparing spectra to the database (red) and the hybrid approach based on constructing spectral dictionaries and fast database lookup (blue). The red lines illustrate that in traditional searches every spectrum should be compared to every peptide in the database with a given parent mass (the running time scales linearly with the database size). The blue lines illustrate that every peptide in the spectral dictionary should be checked for presence in the database (the running time is negligent if the database is pre-processed as a hash table or a suffix tree). The running time of the de novo-based approaches is nearly independent of the database size (it is dominated by the time required to generate the spectral dictionaries). The fast database lookup can be implemented either as exact matching or as error-tolerant lookup (to search for mutations/polymorphisms).

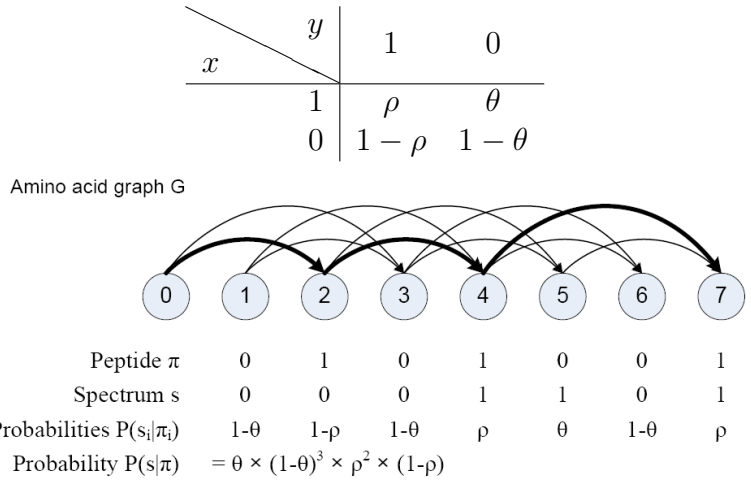


Figure 3.2: *Left*, Probability $P(x|y)$ of a peptide symbol y generating a spectrum symbol x. *Right*, The amino acid graph G for all peptides with parent mass 7 and only two possible "amino acids" A and B with masses 2 and 3 correspondingly. The highlighted path corresponds to the G-peptide 0101001 corresponding to AAB (masses of consecutive amino acid masses are 2,2,3). Two other G-peptides with parent mass 7 are 0100101 (ABA) and 0010101 (BAA). The probability of a spectrum $s = s_1 \dots s_n$ being generated by a peptide $\pi = \pi_1 \dots \pi_n$ is defined as $P(s|\pi) = \Pi_{i=1}^{n}(s_i|\pi_i)$. This is illustrated above with $\pi = 10101001$ and $s = 0001101$ ($P(s = 0001101, \pi = 10101001) = \theta \cdot (1-\theta)^3 \cdot \rho^2 \cdot (1-\rho)$).

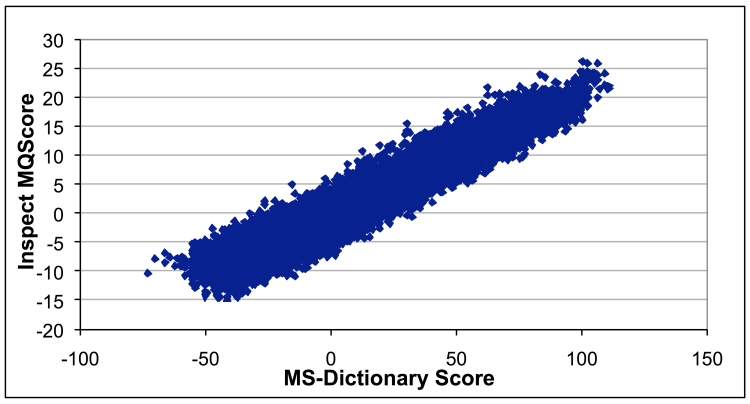
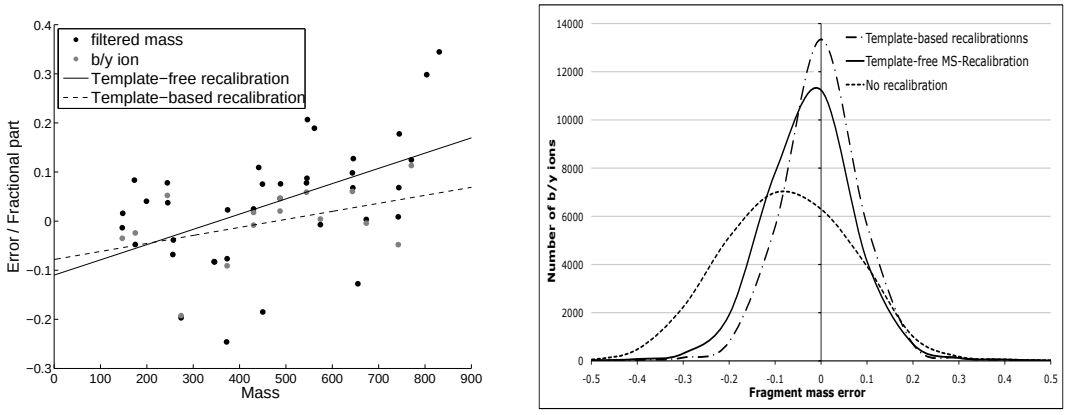
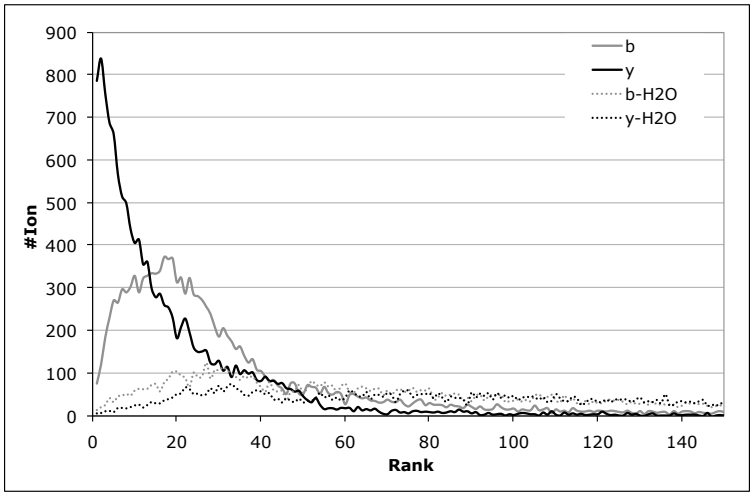


Figure 3.3: Correlation between InsPecT and MS-Dictionary scores computed on randomly selected 50,000 spectra (correlation coefficient is 0.96).

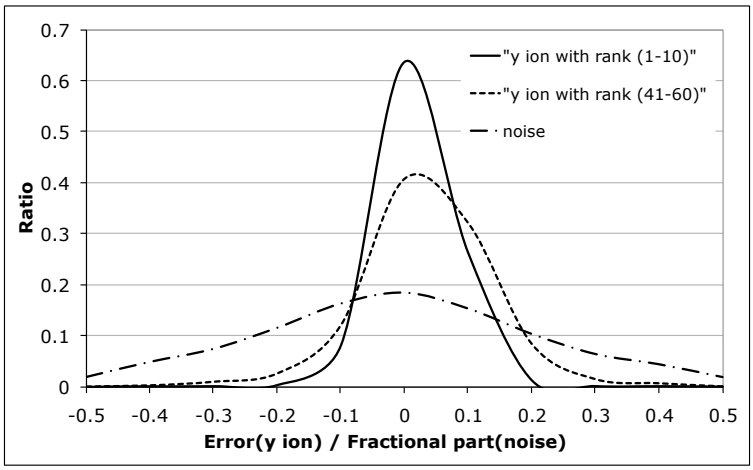


(a) (b)

Figure 3.4: **(a)** Comparison of template-free (solid line) and template-based (dashed line) recalibrations for a single spectrum. Each blue dot represents a 2-D point (m, $Frac(m)$) for a mass m (for every peak in the rescaled and filtered spectrum). Each red dot represents a 2-D point (m, $Error(m)$) for a b- or y-peak with mass m and the difference between the theoretical and experimental mass of the peak equal to $Error(m)$ (for every b- and y- peak in the original spectrum). **(b)** MS-Recalibration performance on 1745 identified spectra of length 10 in the *Shewanella* dataset. The template-based recalibration uses the positions of theoretical b- and y-ions in the spectrum to fit the positions of b- and y-ions in the experimental spectrum using the least-squares fit algorithm. The template-free MS-Recalibration does not require knowledge of the theoretical b- and y- ions. The error distribution for non-calibrated spectra is shown for comparison. The average error is 0.13 before recalbration, 0.07 after MS-Recalibration, and 0.06 after the template-based recalibration. Before recalibration, only 79% of b/y ions are within mass error 0.2 Da as compared to 96% after MS-Recalibration (similar to 98% for the template-based recalibration).

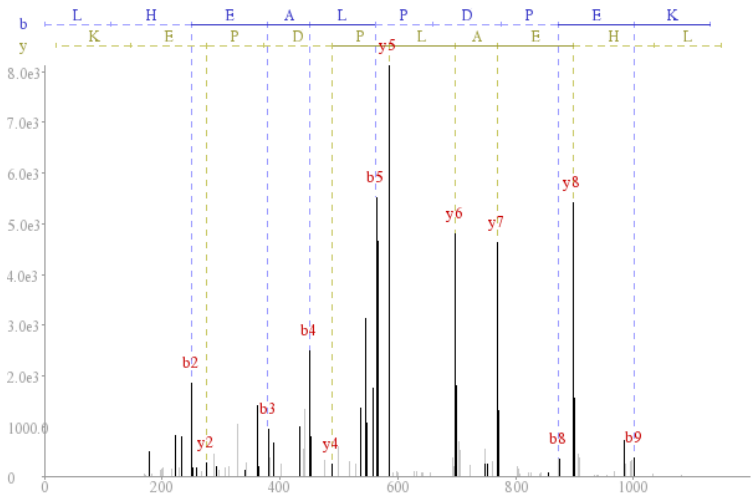


(a)

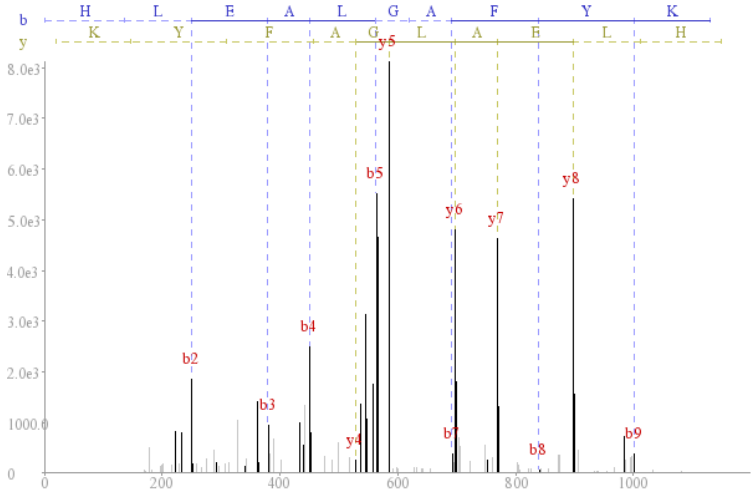


(b)

Figure 3.5: **(a)** Different fragment ions have different rank distributions (statistics is given for all spectra of length 10 from the Shewanella dataset). **(b)** Distributions of mass errors of y peaks depends on their intensity (statistics is given for all spectra of length 10 from the Shewanella dataset). The high intensity peaks (solid curve) tend to have more accurate mass measurements than the lower intensity peaks (dashed curve). The fractional parts of very low intensity peaks (peaks of rank higher than 150) are centered around zero after rescaling (dashed-dot curve).

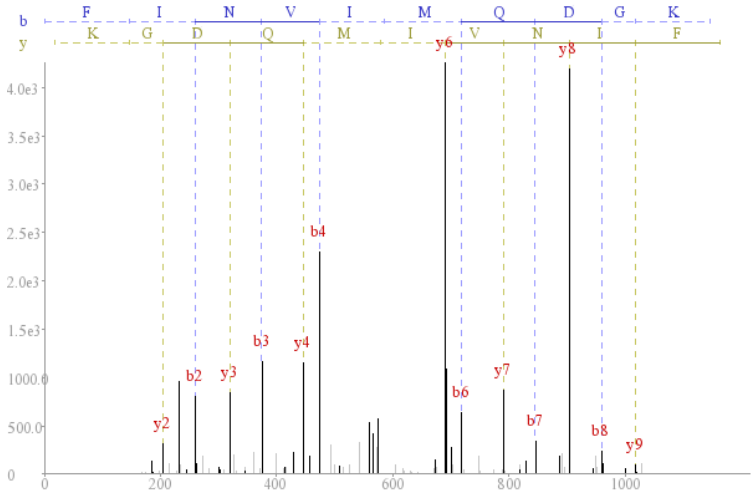


(a)

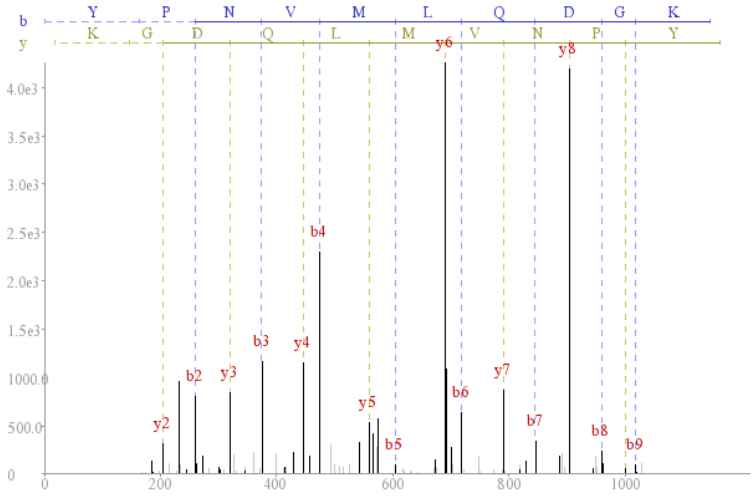


(b)

Figure 3.6: Two optimal de novo interpretations **(a)** LHEALPDPEK and **(b)** HLEALGAFYK for a particular spectrum.



(a)



(b)

Figure 3.7: **(a)** Correct peptide FINVIMQDGK as identified by InsPecT database search and **(b)** YPNVMLQDGK, a de novo reconstruction, for a particular spectrum. The former gets a score of 111 compared to a higher score 123 of the latter.

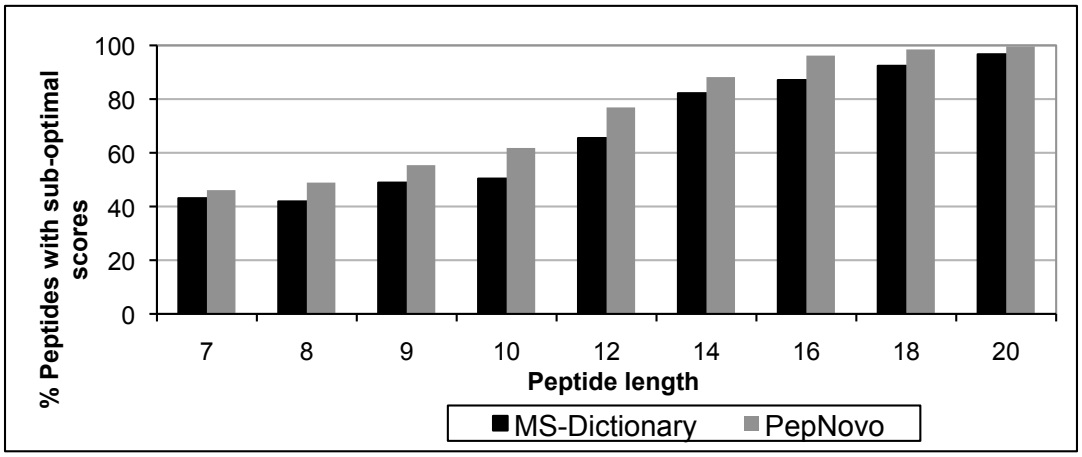


Figure 3.8: Fraction of the spectra for which the correct peptide (as identified by the database search) has a suboptimal de novo score (depending on the length of the spectra). The distribution is shown for MS-Dictionary and PepNovo scoring functions.

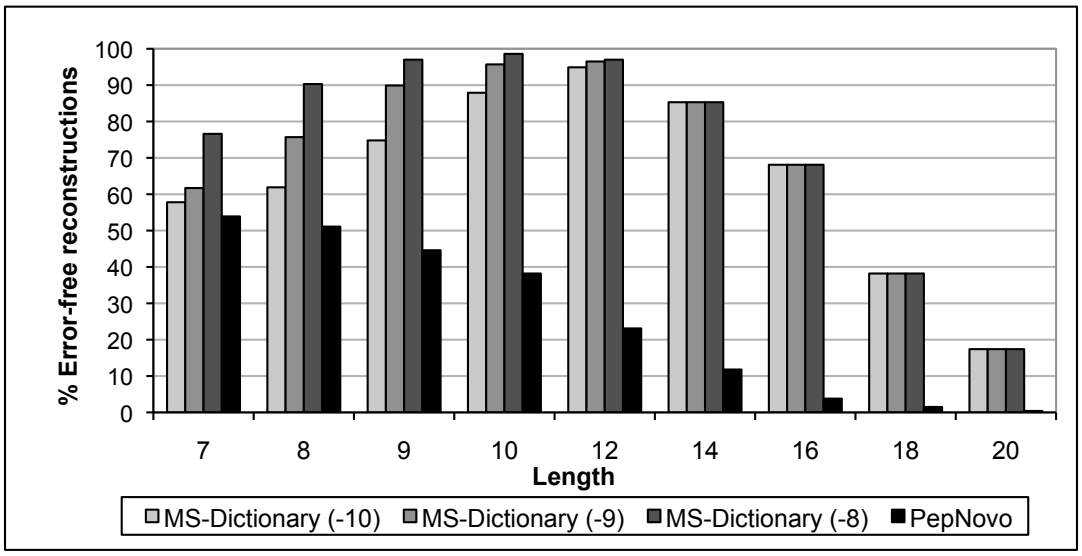


Figure 3.9: MS-Dictionary accuracy as a function of the spectrum length. Percentage of spectra that were correctly reconstructed by MS-Dictionary (i.e. the correct peptide was present in the spectral dictionary) are shown on the y-axis. Accuracies are computed for three different values of $SpectralProbability$, viz. 10^{-10}, 10^{-9} and 10^{-8}. Comparison with PepNovo (counting the percentage of spectra for which PepNovo reconstructs the correct peptide) is shown. As the number of reconstructions for length 14 aa and above is often larger than our allowed limit of 100,000 reconstructions per spectrum, the same set of reconstructions being generated for different $SpectralProbability$ values.

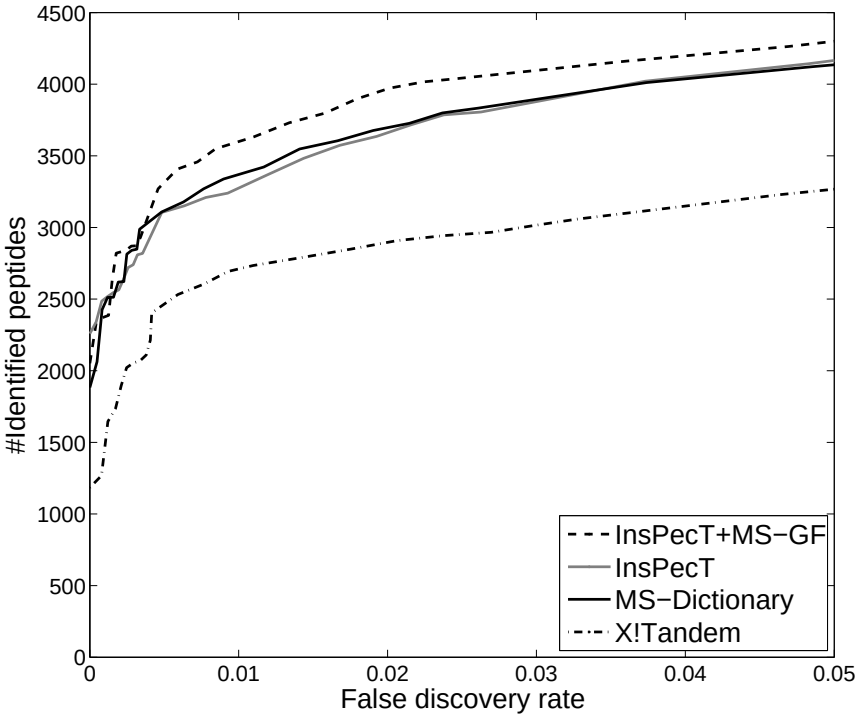
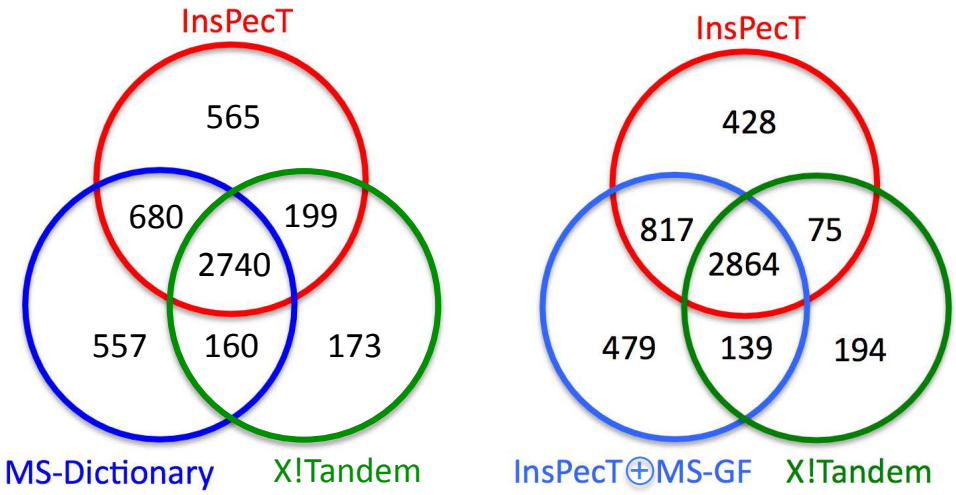


Figure 3.10: Comparison of the number of peptide identifications by various approaches, viz. InsPecT, X!Tandem, MS-Dictionary, and InsPecT $\oplus$ MS-GF. The searches were performed with spectra of charge 2 from the *Shewanella* dataset within the $PrecursorMass$ range from 1100 to 1200 Da. The curves display the number of peptide identifications for different score thresholds (corresponding to different false discovery rates).



(a) (b)

Figure 3.11: Venn diagram showing the overlap between peptides identified by different approaches at 5% false discovery rate. **(a)** Overlap between InsPecT, X!Tandem and MS-Dictionary. **(b)** Overlap between InsPecT, X!Tandem and InsPecT $\oplus$ MS-GF.

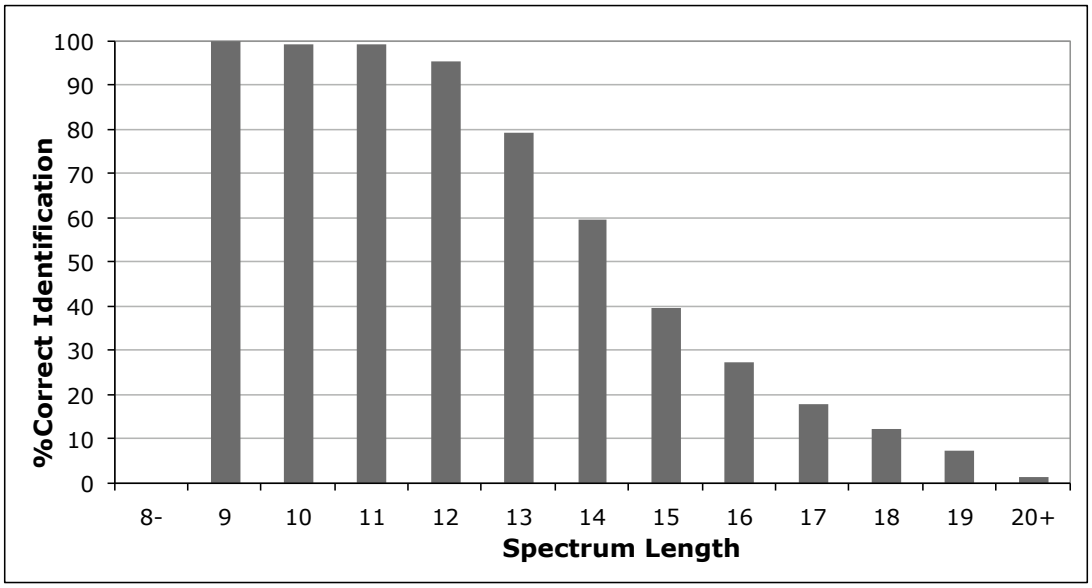


Figure 3.12: The percentage of peptides identified by MS-Dictionary in the translated human genome as compared to all peptides identifies in searches of human protein database. Spectral dictionaries were generated for the 21,635 selected spectra from HEK293 dataset and searched against the translated human genome. For each spectrum, if correct peptide is contained in the dictionary of the spectrum, we regarded the spectrum as identified.

Chapter 4

Spectral Profiles and Their Applications for de Novo Peptide Sequencing

4.1 Introduction

Recent advances in de novo peptide sequencing have enabled tag-based peptide identification tools (e.g., Inspect [25] and Paragon [47]) that are orders of magnitude faster than traditional MS/MS database search approaches (e.g., Sequest [24] and Mascot [5]). However, reliable full-length de novo peptide sequencing remains an elusive goal, and even the most accurate de novo tools correctly reconstruct only 30–45% of peptides [35]. We argue that accurate full length de novo peptide sequencing may be an unattainable goal for many spectra since they do not provide enough information to disambiguate between correct and incorrect reconstructions. Spectra often have variable local quality (along the peptide length) making some regions not amenable to de novo sequencing. For example, spectra of peptides D**GE**AAENTDAQK and D**SV**AAENTDAQK are very similar making it nearly impossible to reliably reconstruct these peptides de novo (the combined mass of G and E is close to the combined mass of S and V). In such cases, it makes more sense to reconstruct a *gapped* peptide D[**186**]AAENTDAQK rather than a

contiguous peptide. While gapped peptides are less informative than full-length peptides, we argue that there is little difference between these two representations. Indeed, in most applications, de novo peptide sequencing is not the final goal in analyzing a spectrum but rather a prelude to error-tolerant database searches and other applications like metaproteomics [91, 59, 92, 93]. We argue that *long* gapped peptides are nearly as good for such applications as full-length de novo reconstructions. For example, the gapped peptide D[**186**]AAENTDAQK has 9 continuous amino acids and thus, for all practical applications, is at least as useful as any peptide of length 9 (or length 11 if one counts D and [**186**] as separate "letters"). Since most mass-spectrometrists view peptides of length 9 as useful as peptides of length 12, generating sufficiently long gapped peptides is nearly as useful as generating full-length reconstructions (the full length of D[**186**]AAENTDAQK is 12).

In this chapter we introduce the notion of a *spectral profile* (Fig. 4.1) that enables accurate de novo sequencing of gapped peptides and reveals the variable spectral quality along the peptide length. For example, for peptides of length 11-12, our MS-Profile tool correctly reconstructs 65% of gapped peptides as compared to 46%, 28% and 26% correct reconstructions of full or truncated full-length peptides by PepNovo+ [35, 94], MS-Dictionary [3], and PEAKS [38]. Gapped peptides occupy a niche between peptide sequence tags (that in most applications are limited to tags of length 3) and full-length reconstructions: they are nearly as accurate as short tags and, at the same time, typically have a unique match in the protein database. E.g., for peptides of length 12, the average length of gapped reconstructions is 8.9, typically resulting in a single hit even when searching against the largest databases used in proteomics today.[1]

A spectral profile is a novel representation of tandem mass spectra with "intensities" of all masses varying from 0 to 1. Every peptide of length n defines n *prefix* masses representing masses of the first i amino acids (for $1 \leq i \leq n$). The spectral profile at mass x is the proportion of peptides with prefix mass x among all

[1]We define the length of the gapped peptide as the number of masses *and* amino acids describing the peptide. For example, the length of the gapped peptide [**186**]DK[**246**]FK is 6, while the length of a 3-aa long peptide sequence tag [**307**]GTP[**421**] is 5.

high-scoring interpretations of the spectrum. Thus, the spectral profile compactly represents information about *all* high-scoring de novo reconstructions (*spectral dictionary*) even if there are billions of such reconstructions (see [3]). Spectral profiles are conceptually similar to the motif profiles [95] that are used in various areas of bioinformatics (e.g., in regulatory genomics). While motif profiles in regulatory genomics compactly represent all known binding sites of a transcription factor, a spectral profile compactly represents all high-scoring de novo reconstructions of an MS/MS spectrum. However, while motif profiles represent the center of gravity of *known* motifs, spectral profiles represent the center of gravity of *unknown* high-scoring de novo reconstructions (spectral dictionaries). This makes computing spectral profiles challenging since in many cases spectral dictionaries cannot be explicitly generated [3]. This chapter extends Chapter 3 by showing how to compute spectral profile of *any* spectrum without explicitly generating its spectral dictionary. We further show how to use spectral profiles for generating reliable gapped peptides.

The difficult challenge in de novo spectral interpretations is how to figure out which ion type every peak represents (e.g., how to distinguish b-series peaks from y-series peaks) and how to analyze the widely varying intensities in a single probabilistic framework. The spectral profile collapses all possible ion type interpretations and varying intensities into a single ion type (b-ion) with rigorously defined probability. In difference from real MS/MS spectra (that contain peaks corresponding to b- and y-ions, various neutral losses, etc.), spectral profiles only represent (putative) b-ions. In some sense, spectral profiles represent a trade-off between (hard-to-interpret but compact) real spectra and (easy-to-interpret but huge) spectral dictionaries. We emphasize that spectral profiles are different from "scored spectra" (e.g., sequence spectra [80, 62] or prefix residue mass spectra [25]) that are commonly used for de novo sequencing and MS/MS database searches. While profile probabilities are *global* (i.e., they take into account complex dependencies between all peaks in the spectrum), scored spectra take into account only a few *local* satellite peaks explaining a given mass.

Similar to the diverse applications of motif profiles, spectral profiles have a

multitude of applications that we describe below. Fig. 4.2 illustrates recently implemented alternative approaches to peptide identification: peptide sequence tag approaches [25, 47, 96, 97] and full length de novo reconstruction approaches [59, 92, 3, 56] (see also lookup-peak approach [46]). While these approaches significantly speed up conventional peptide identification tools, each of them presents certain challenges, leading to deteriorating performance on long (15+ aa) peptides. Most of these approaches do not automatically adjust to varying spectral qualities or different peptide lengths. For example, InsPecT generates the same number of tags for every spectrum while a more sensible approach would be to generate a larger number of tags for long peptides (tag generation deteriorates for longer peptides) or for low-quality spectra. While MS-Dictionary [3] generates an adaptive but large number of full length reconstructions (for both high- and low-quality spectra), dictionaries of spectra of long peptides may become so large that their generation becomes impractical. To overcome this problem, we show how to quickly construct spectral profiles of even huge dictionaries without explicitly generating them.

MS-Profile currently works in two modes (Fig. 4.3). In the first mode, the input is an MS/MS spectrum and a spectral probability threshold (described below) and the output is a spectral profile. In the second mode, the constructed spectral profile, in addition to a de novo reconstruction, and a $MinProbability$ threshold (described below) serve as an input, and the output is a gapped peptide. MS-Profile in the second mode represents a new de novo peptide sequencing tool that improves accuracy of de novo reconstructions produced by other tools (e.g., PepNovo+, PEAKS, or MS-Dictionary). In particular, it generates gapped peptides that can be used for mutation-tolerant database searches and speed up existing database search tools. MS-Profile is available both as open source software and as a web server.

4.2 Methods

What is the spectral profile? For the sake of simplicity, we will first

introduce the notion of a spectral profile under the assumption that amino acid masses are integer[2]. Given a peptide $p_1 \dots p_n$, we define its *prefix masses* as a series

$$mass(p_1), mass(p_1) + mass(p_2), \dots, \sum_{j=1}^{i} mass(p_j), \dots, \sum_{j=1}^{n} mass(p_j)$$

where $\sum_{j=1}^{n} mass(p_j) = k$ is defined as the *parent mass*. We further represent the peptide $p_1, \dots p_n$ as a k-mer boolean vector $P = x_1 \dots x_k$, where $x_t = 1$, if t represents a prefix mass, and $x_t = 0$, otherwise (see [3, 29, 79] for applications of boolean spectra and peptides). Given a set of boolean peptides $Dictionary = \{P_1, \dots, P_m\}$, we define the spectral profile as simply the center of gravity of all peptides (boolean vectors) in the set, i.e., $Profile(Dictionary) = \frac{1}{m} \sum_{j=1}^{m} P_j$. This definition assumes that all peptides in the *Dictionary* are equally likely.

Kim et al., 2008 [3] introduced the notion of *spectral dictionary* and described an MS-Dictionary approach to peptide identification. Given a spectrum *Spectrum* and a score threshold *Threshold*, $Dictionary(Spectrum, Threshold)$ is defined as the set of all peptides with scores above the *Threshold*. We define the spectral profile $Profile(Spectrum, Threshold)$ as $Profile(Dictionary(Spectrum, Threshold))$.

When the *Dictionary* is explicitly given, computing $Profile(Dictionary)$ amounts to computing the center of gravity of k-dimensional boolean vectors from *Dictionary*. While MS-Dictionary [3] is capable of quickly generating spectral dictionaries for short peptides (less than 15 aa), the spectral dictionaries of spectra of long peptides are so large (even for sensible choices of *Threshold*) that MS-Dictionary becomes impractical. For example, for a typical spectrum of a 15-aa long peptide, the spectral dictionary consists of $\approx 4 \cdot 10^9$ high-scoring peptides that would typically result in statistically significant database hits [3]. Below we show how to quickly generate spectral profiles of such huge dictionaries without explicitly generating the dictionary. MS-Profile takes only ≈ 0.2 seconds to generate the spectral profiles even for spectra of long peptides. Thus MS-Profile bypasses the

[2]One can always adjust the "granularity" of mass measurements (e.g., by multiplying all masses by 1000 in case of accurate mass measurements) and to safely assume that the masses of amino acids become integer after this transformation.

need to explicitly generate large spectral dictionaries that limited applications of MS-Dictionary in the case of long peptides.

Computing spectral profiles. The transformation of spectra into spectral profiles can be done efficiently by the *forward-backward* dynamic programming algorithm [98]. For the sake of simplicity, we first represent a spectrum with parent mass k as a *boolean* spectrum $S = s_1 \dots s_k$, where $s_i = 1$ if there is a peak at mass i in the spectrum, and $s_i = 0$, otherwise. This representation assumes that spectra are discretized and all masses are integers. Below we use the term *mass* of peptide/spectra to refer to the dimension of the corresponding vectors (parent mass k). The score (denoted as $Score(P, S)$) between a boolean peptide $P = p_1 \dots p_k$ and a boolean spectrum $S = s_1 \dots s_k$ (of the same mass) is defined as $\sum_{j=1}^{k} p_j \cdot s_j$. When peptide P and spectrum S differ in mass, we define $Score(P, S)$ as $-\infty$.

Define S_i^{prefix} as $s_1, \dots s_i$ and S_i^{suffix} as $s_{k-i+1}, \dots s_k$. Given a spectrum $S = s_1 \dots s_k$, define $\mathcal{P}^{prefix}(i, t)$ as the set of all boolean peptides $P \in \mathcal{P}^{prefix}(i, t)$ with length i and $Score(P, S_i^{prefix}) = t$. Let $x_{fwd}(i, t)$ be the size of $\mathcal{P}^{prefix}(i, t)$. As shown in [78], $x_{fwd}(i, t)$ can be computed using the *forward* dynamic programming:

$$x_{fwd}(i,t) = \sum_{\text{all amino acids } a} x_{fwd}(i - mass(a), t - s_i)$$

We initialize $x_{fwd}(0, 0) = 1$, $x_{fwd}(0, t) = 0$ for $t > 0$, and set $x_{fwd}(i, t) = 0$ for negative i.

Given a spectrum $S = s_1 \dots s_k$, define $\mathcal{P}^{suffix}(i, t)$ as the set of all boolean peptides $P \in \mathcal{P}^{suffix}(i, t)$ with length $k-i$ and $Score(P, S_{k-i}^{suffix}) = t$. Let $x_{bwd}(i, t)$ be the size of $\mathcal{P}^{suffix}(i, t)$. The variable $x_{bwd}(i, t)$ can be computed using the *reverse* dynamic programming:

$$x_{bwd}(i,t) = \sum_{\text{all amino acids } a} x_{bwd}(i + mass(a), t - s_{i+mass(a)})$$

We initialize $x_{bwd}(k, 0) = 1$, $x_{bwd}(k, t) = 0$ for $t > 0$, and set $x_{bwd}(i, t) = 0$ for $i > k$.

Given a score threshold $Threshold$ to generate a $Dictionary$, it is easy to

see that the size of the *Dictionary* can be computed as follows:

$$|Dictionary| = \sum_{t>Threshold} x_{fwd}(k,t)$$

Below we demonstrate that $Profile(Spectrum, Threshold) = f_1 \dots f_k$ can be computed using the *forward-backward* algorithm:

$$f_i = \frac{1}{|Dictionary|} \sum_{t+t'>Threshold} x_{fwd}(i,t) \cdot x_{bwd}(i,t').$$

Indeed,

$$f_i = \frac{\#\text{peptides } x_1 \dots x_k \in Dictionary \text{ with } x_i = 1}{|Dictionary|}.$$

Every peptide in *Dictionary* with $x_i = 1$ can be decomposed into $Pref = x_1 \dots x_i$ and $Suff = x_{i+1} \dots x_k$ peptides. Since all peptides in the *Dictionary* score above $Threshold$, $Score(Pref, s_1 \dots s_i) + Score(Suff, s_{i+1} \dots s_k) > Threshold$. Thus, the number of such peptides is given by $\sum_{t+t'>Threshold} x_{fwd}(i,t) \cdot x_{bwd}(i,t')$. Conversely, concatenation of two arbitrary peptides $Pref = x_1 \dots x_i$ and $Suff = x_{i+1} \dots x_k$ contributes to the *Dictionary* as long as $Score(Pref, s_1 \dots s_i) + Score($ $Suff, s_{i+1} \dots s_k) > Threshold$. Since the number of such concatenations with $x_i = 1$ is given by $\sum_{t+t'>Threshold} x_{fwd}(i,t) \cdot x_{bwd}(i,t')$, f_i can be computed by the forward-backward algorithm as described above.

Figure 4.4 illustrates computing a spectral profile. In practice, we compute spectral profiles for a fixed *spectral probability* [78] (rather than for a fixed score threshold). The spectral probability of a *Peptide-Spectrum Match (PSM)* is defined as the total probability of all peptides with scores exceeding the score of the PSM.[3] One can also define a spectral probabilty depending on a score $Threshold$ as the total probability of all peptides with scores above $Threshold$ (the total probability of all peptides in the corresponding spectral dictionary). Given a spectral probability p, one can approximate the E-value as $p \cdot DatabaseSize$. See [3, 78] for the background on spectral probabilities and spectral dictionaries. For each spectrum, MS-Profile dynamically sets $Threshold$ as the minimum score s such that

[3]The probability of a peptide is defined as the product of probabilities of its amino acids. Amino acid probabilities are pre-defined depending on the frequencies of amino acids in a protein database [78].

the spectral probability of the reconstructions with scores above s doesn't exceed a predefined spectral probability (e.g., 10^{-8}) and computes the spectral profile. For example, the spectral profile in Fig. 4.1 was computed for spectral probability 10^{-8}. The spectral profile remains stable for a range of spectral probabilities.

Note that the simple boolean model for scoring peptide-spectrum matches can easily be extended to more complicated models without any algorithmic changes. Indeed, MS-Profile uses MS-Dictionary's scoring model [78] that considers various features such as ion types, peak intensities and mass errors.

4.3 Results

Dataset. We used the Standard Protein Mix database consisting of 1.1 million spectra generated from 18 proteins using 8 different mass spectrometers [99]. For this study, we considered only the charge 2 spectra generated by Thermo Electron LTQ where 1388 peptides of length between 7 and 20 are reliably identified with false discovery rate 2.5% using Sequest [24] and PeptideProphet [11] in the search against the *Haemophilus influenzae* database appended with sequences of the 18 proteins (567,460 residues). Although this chapter focuses on doubly charged spectra, MS-Profile can also be applied to MS/MS spectra of higher charges as long as additive scoring model for highly charged MS/MS spectra is available.

For each peptide, we randomly selected one representative spectrum and formed a dataset of 1388 PSMs grouped by the length of their peptide identifications. To avoid computational artifacts introduced by errors in the parent mass, the parent masses of the spectra is corrected according to the Sequest identifications. Below, we refer to this dataset as the *Standard dataset*. Throughout this chapter, we measure *accuracy* of a de novo sequencing tool as the percentage of spectra with error-free reconstructions among all spectra in the Standard dataset.

Fig. 4.5 shows the distribution of spectral probabilities (false positive rates) of the Standard dataset. Most PSMs (91%) have spectral probabilities lower than 10^{-8}. We used 10^{-8} as the spectral probability threshold to generate spectral

profiles.

Table 4.1 shows the results of de novo peptide sequencing of the Standard dataset with PEAKS, PepNovo+, and MS-Dictionary. PEAKS and MS-Dictionary correctly reconstructed peptides for less than 30% of the spectra and the accuracy of both tools greatly deteriorates as the peptide length increases. PepNovo+ reported shorter de novo reconstructions (especially for spectra of long peptides) by allowing gaps in the start and the end of the peptides, resulting better accuracy than the other tools. Below we show that MS-Profile improves the accuracy of these tools at the cost of a small reduces in the length of reconstructed peptides.

De novo sequencing of gapped peptides. De novo peptide sequencing algorithms usually correctly recover some amino acids within a peptide and misinterpret others. The key challenge is to figure out which portions of the peptide are reconstructed incorrectly and to limit reconstructions to highly accurate portions. Gapped peptide reconstruction addresses this challenge by reporting only reliably reconstructed regions of the peptide.

Given a $Peptide = x_1 \dots x_k$, a $Profile = f_1 \dots, f_k$, and a parameter $MinProbability$, we define $GappedPeptide(Peptide, Profile, MinProbability) = g_1 \dots g_k$ as $g_i = x_i$ if $f_i \geq MinProbabilty$ and $g_i = 0$ otherwise. Fig. 4.1 shows a spectral profile for the spectrum of peptide STVAGESGSADTVR and (incorrect) de novo reconstruction SSLAGESGSADTVR. One can notice that while profile values for most prefix masses in STVAGESGSADTVR are relatively high (0.207, **0.084**, 0.475, 0.518, 0.310, 0.522, 0.791, 0.718, 0.730, 0.709, 0.323, 0.149, 0.353), the profile value for one prefix mass falls below 0.1. This low profile value points to an unreliable portion of the reconstruction. Converting peptide STVAGESGSADTVR into a gapped peptide (with $MinProbability = 0.1$) results in a (correct) gapped peptide S[**200**]AGESGSADTVR. Increasing $MinProbability$ to 0.2 results in a shorter gapped peptide S[**200**]AGESGSAD[**200**]R.

MS-Profile generates gapped peptides as follows. For each spectrum, it first constructs the spectral profile and generates optimal de novo reconstructions by backtracking its forward matrix. Indeed, since MS-Profile uses the MS-Dictionary scoring [3], the reconstructions are the same as reconstructions generated by MS-

Dictionary. Both PEAKS and MS-Dictionary may generate (a small number of) multiple optimal de novo reconstructions, and we first convert them into a single *consensus* reconstruction. For example, the set of reconstructions YWA**GE**LTR, YWA**SV**LTR, YWA**VS**LTR, YWA**EG**LTR will be converted into a single consensus reconstruction YWA[**186**]LTR by retaining only the prefix masses present in all reconstructions. Next, MS-Profile discards all prefix masses in the consensus reconstruction whose corresponding profile values are below $MinProbability$ as described above. The remaining prefix masses represent the gapped peptide generated by applying MS-Profile to MS-Dictionary (referred to as MS-Profile(MS-Dictionary)). Fig. 4.6 compares the accuracy of de novo reconstructions generated by MS-Dictionary and the gapped peptide generated by MS-Profile(MS-Dictionary). is defined as the percentage of the error-free reconstructions among all recontructions for the Standard dataset. Applying MS-Profile increases the percent of correct reconstructions from 28% to 42% while decreasing the average length of reconstructions from 12.8 to 9.1 amino acids when $MinProbability = 0.1$. We remark that the Standard dataset contains some low-quality spectra that are nearly impossible to reconstruct in de novo fashion. One can increase the accuracy by increasing the $MinProbability$ threshold. For example, when $MinProbability = 0.2$, the accuracy increases to 50% while the average length of gapped peptide decreases to 7.9. When $MinProbability = 0.3$, the accuracy increases to 54% while the average length of gapped peptide becomes 7.2.

PepNovo+ and PEAKS represent some of the most accurate de novo peptide sequencing tools. MS-Profile can be used to convert PepNovo+ and PEAKS reconstructions into gapped peptides resulting in MS-Profile(PepNovo+) and MS-Profile(PEAKS) tools. Applying MS-Profile to PepNovo+ increases the percent of correct reconstructions from 46% to 65% while decreasing the average length of reconstructions from 11.0 to 8.9 amino acids ($MinProbability = 0.1$). Applying MS-Profile to PEAKS increases the percent of correct reconstructions from 26% to 48% while decreasing the average length of reconstructions from 12.6 to 9.2 amino acids ($MinProbability = 0.1$). Although gapped peptides generated by MS-Profile(PepNovo+) and MS-Profile(PEAKS) are shorter than PepNovo+ and

PEAKS reconstructions, they are still long enough to uniquely identify most peptide even in large protein databases. Fig. 4.7 compares the accuracy and lengths of PepNovo+, PEAKS, MS-Profile(PepNovo+), and MS-Profile(PEAKS) reconstructions.

PepNovo+ allows users to generate up to 2000 reconstructions per spectrum. When multiple reconstructions are generated, the probability of at least one of them being correct increases. For each reconstruction, we generate a gapped peptide using MS-Profile(PepNovo+). Since different PepNovo+ reconstructions may correspond to the same gapped peptide, the number of gapped peptides generated by MS-Profile(PepNovo+) is typically smaller than the original number of PepNovo+'s reconstructions. While the number of gapped peptides generated by MS-Profile(PepNovo+) is 3-15 times smaller than the number of PepNovo+'s reconstructions, the length of the reconstructed gapped peptides is typically sufficient to ensure a unique database hit. Fig. 4.8 compares accuracy and length of peptides and gapped peptides generated by PepNovo+ and MS-Profile(PepNovo+) for the top 100 and the top 1000 reconstructions. Again, MS-Profile(PepNovo+) outperforms PepNovo+ while generating much smaller numbers of gapped peptides.

The improved performance of MS-Profile(PepNovo+) in generating gapped peptides suggests that it can be used for database filtration in the same way as peptide sequence tags in InsPecT [25]. For the Standard dataset, we ran InsPecT to generate 1, 10 and 25 tags of length 3 and 4 and measured for how many spectra InsPecT generates at least one correct tag (Fig. 4.9). The same number of gapped peptides is also generated by MS-Profile(PepNovo+). It turned out that the best gapped peptide is longer and more accurate than the best tag of length 3 (the gapped peptide is correct for 65% of spectra while the best 3 aa long tag is correct for 44% of spectra). Also, top 10 and 25 gapped peptides are roughly as accurate as the same number of tags of length 3. For 83% of spectra, at least one of top 10 gapped peptides are correct while for 80% of spectra, at least one of top 10 tags of 3 aa are correct. For 86% of spectra, at least one of top 25 gapped peptides are correct while for 88% of spectra, at least one of top 25 tags of 3 aa are correct. This is surprising, since gapped peptides generated

by MS-Profile(PepNovo+) represent a much better filter for database search than InsPecT tags. To test the filtering efficiency, we matched each spectrum's top gapped peptide and its top 3 aa tag against the Swiss-Prot database (Release 56.4, 145 million residues) counting the number of false matches to the database. While 90% of gapped peptides have no false matches, only 29% of tags have no false match. The average number of false matches is 1.6 for gapped peptides, fifty times smaller than 80.3 false tag matches on average. The average number of false matches is a key parameter in filtration-based MS/MS searches since it is roughly proportional to the time required for peptide identification [25]. Therefore, fifty-fold reduction in the number of false matches can potentially translate into fifty-fold speed-up as compared to (already fast) InsPecT. The contrast between gapped peptides and tags is particularly pronounced in searches against very large databases like proteogenomic six-frame translation searches of the repeat-masked human genome of size 2.7 billion residues [3]. Gapped peptides longer than 8 aa (63% of spectra in the Standard dataset) are expected to have only 0.24 false matches in this database while 3 aa tags are expected to have 1400 false matches on average.

This comparison suggests that MS-Profile can significantly improve on previous filtration approaches to MS/MS database searches. In difference from peptide sequence tags (that typically have many false hits in a database), gapped peptides typically have few false hits (if any) thus speeding up the database searches. We comment that use of gapped seeds in traditional BLAST-like genomics searches is well studied [100].

Evaluating spectral profile probabilities. Some de novo sequencing programs output the reliability of predicted amino acids. For example, PepNovo+ defines features that reflect the reliability of each predicted amino acid and converts the feature vectors into probabilities [57]. PEAKS recently added a similar function that computes the reliability of an amino acid a by locally permuting the reconstruction around a, computing the score difference between the original and permuted reconstructions, and using the pre-learned distribution of the difference to assign the reliability of a [101]. MS-Profile differs from these tools since instead

of learning, it rigorously computes a probability that a prefix mass is present in a high-scoring de novo peptide reconstruction.

We show that the spectral profile probabilities approximate the empirical accuracy of the prefix mass (represented by the profile peak) being correct. To compute the accuracy of the profile value p (for $p = 0.1, 0.2, \ldots 0.9, 1.0$), we bundled all the profile peaks with values between $p - 0.05$ and $p + 0.05$ and measured the fraction of correct peaks among them. If the empirically computed fraction of correct peaks of the profile value p is close to p then our estimate of profile probabilities is unbiased. Fig. 4.10 shows that it is indeed the case: the empirical accuracy of the profile peaks with probability p is slightly above p. The slightly higher empirical accuracy (as compared to profile values) is likely a consequence of using the same spectral probability threshold 10^{-8} for all spectra while in reality most PSMs have much lower spectral probabilities (Fig. 4.5).

4.4 Discussion

While peptide sequence tags were first proposed in 1994 [60], it took 10 years for this idea to become an integral part of the new generation of fast MS/MS database search tools [25, 47]. It took such a long time because a seemingly simple problem of generating *accurate* sequence tags turned out to be more difficult than originally thought. We demonstrated that gapped peptides occupy an important niche between long but inaccurate full-length peptide reconstructions and short but more accurate peptide sequence tags. This niche provides certain advantages since gapped peptides represent a more stringent filter that may enable very fast MS/MS database searches that in many cases will amount to a simple look-up in a database. Spectral profiles reveal poor quality spectra (or poor quality regions within long peptides) that other methods have difficulties analyzing. MS-Profile follows a different route to error-tolerant peptide identifications than OpenSea [59] and SPIDER [92]. Instead of trying to generate (unreliable) full-length reconstructions and *approximately* matching them against the database, MS-Profile generates reliable gapped peptides and matches them against the database *exactly*.

Some de novo sequencing tools such as Lutefisk [32], PEAKS [38] and PepNovo+ [94] can generate gapped peptides typically trimming the full length peptides in the beginning/end. Even when internal gaps are allowed (Lutefisk and PEAKS), they are limited to gaps of 2 aa or shorter. For long peptides where multiple consecutive peaks are missing, it is hard to generate correct gapped peptides when only short gaps are allowed. On the other hand, PepNovo+ improves on these tools by allowing long gaps in the start/end of the peptides. As a result, PepNovo+ has a tendency to generate incorrect solutions when it tries to reconstruct all amino acids in the middle. To the best of our knowledge, MS-Profile is the only program that allows both short and long gaps regardless of the position. Secondly, MS-Profile can convert any de novo reconstructions into gapped peptides thus making it a useful addition to various de novo peptide sequencing tools.

4.5 Acknowledgements

This chapter, in full, was published as "Spectral profiles: A Novel Representation of Tandem Mass Spectra and Its Applications for De Novo Peptide Sequencing and Identification". S. Kim, N. Bandeira, and P. Pevzner. *Molecular & Cellular Proteomics*, vol. 8, no. 6, pp. 1391-14009, 2009. The dissertation author was the primary author of this paper.

Table 4.1: Accuracy and average length of PEAKS, PepNovo+ and MS-Dictionary for the Standard dataset. PEAKS (online 2.0), PepNovo+ (release 20080724) and MS-Dictionary (release 20071107) were run with parent mass and fragment mass tolerance 0.5 Da, fixed modification of C+57, no optional modifications and without any enzyme preference. In difference from PEAKS and MS-Dictionary, PepNovo+ allows gaps at the start/end of peptides thus giving PepNovo+ significant leverage when it comes to the reported accuracy of reconstruction. Although MS-Dictionary is designed for generating spectral dictionaries (rather than ensuring that the correct reconstruction has the top score), it can be used in de novo mode as well (it has slightly higher accuracy than PEAKS while generating slightly longer peptides). PEAKS and MS-Dictionary have a tendency to output de novo reconstructions that are longer than the correct peptides (e.g., for petides of length 11-12, the average length of PEAKS and MS-Dictionary reconstructions is 12.1 and 12.2). Accuracy of each tool is defined as the percentage of the error-free reconstructions among all recontructions for the Standard dataset.

Peptide Length	PEAKS		PepNovo+		MS-Dictionary	
	Accuracy	Length	Accuracy	Length	Accuracy	Length
7-8	0.59	8.1	0.65	7.8	0.54	7.8
9-10	0.33	10.0	0.54	9.3	0.40	9.9
11-12	0.27	12.1	0.47	10.4	0.31	12.2
13-14	0.12	14.1	0.36	12.0	0.14	14.2
15-16	0.10	16.0	0.34	13.2	0.12	16.1
17-18	0.07	16.9	0.29	14.1	0.02	18.3
19-20	0.05	19.0	0.38	14.0	0.04	20.8
Total	0.26	12.6	0.46	11.0	0.28	12.8

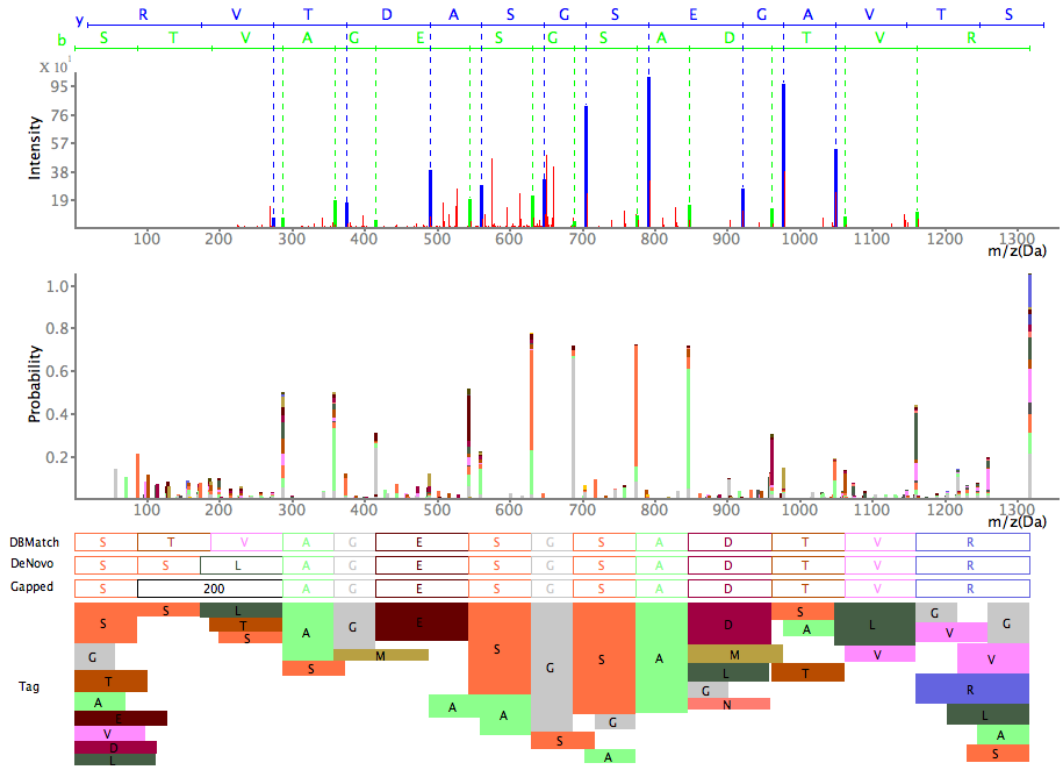


Figure 4.1: An example of the spectral profile. (Top) An MS/MS spectrum of the peptide STVAGESGSADTVR with b- and y-peaks painted green and blue, respectively. (Middle) The spectral profile of the above spectrum. The overall height of each peak represents the probability of the peak being a correct prefix mass. Each peak is represented as a multi-colored bar where various colors (sub-peaks stacked on top of each other) correspond to various amino acids (amino acids are color-coded). Similarly to a motif profile, the height of each colored sub-peak (corresponding to an amino acid X) represents the probability of a prefix with terminal amino acid X ending at the given mass position. (Bottom) The database match (DBMatch), full-length de novo reconstruction (DeNovo) and gapped peptide (Gapped) of the spectrum at the top panel. The painted rectangles represent the tags of length 1 ending at each position of the de novo reconstruction: the width of each rectangle corresponds to the mass of the amino acid and the height corresponds to the probability of the length 1 tag being correct. While the De-Novo reconstruction is incorrect, the Gapped reconstruction (generated using the spectral profile) is correct. The consecutive amino acids S and L are represented as a 200 Da gap since the value of the spectral profile at the position separating S and L is low.

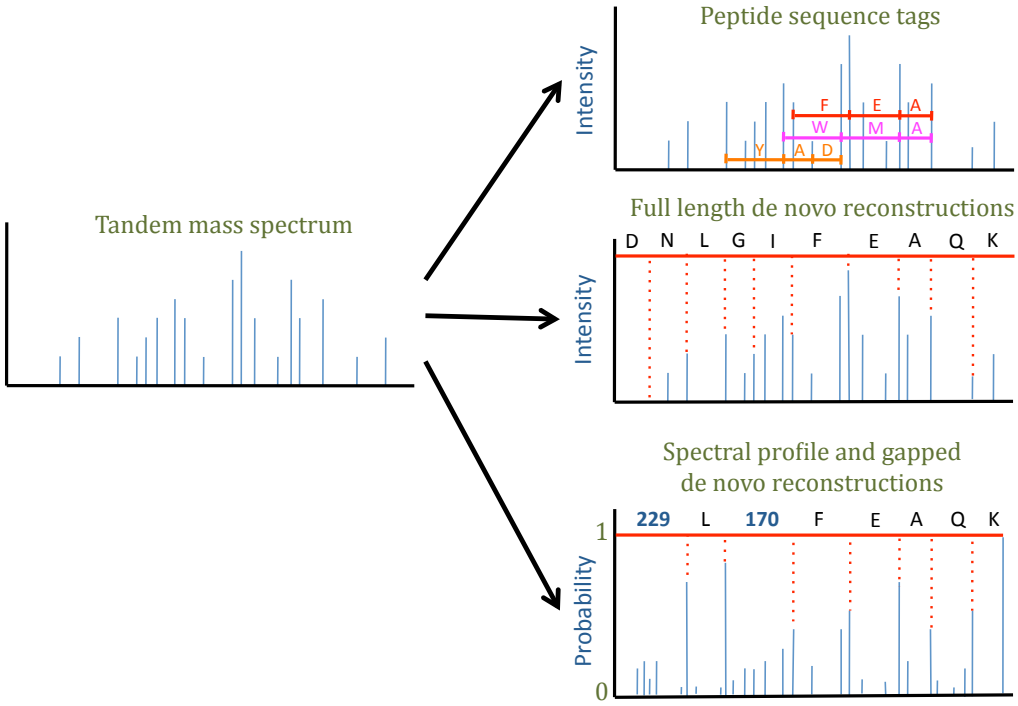


Figure 4.2: Various filtering approaches to peptide identifications. The tag-based approach (e.g., InsPecT [25], Paragon [47]) extracts short (usually length 3) peptide sequence tags and filters databases by considering only peptides that match tags. Full length de novo approaches either reconstruct a single full-length peptide and find sequence matches (e.g., MS-BLAST [91], OpenSea [59] and SPIDER [92]) or generate multiple full length reconstructions and find sequence matches to the protein database (RAId [56] and MS-Dictionary [3]). Spectral profiles represent an alternative approach to peptide identification generating gapped peptides and matching them to the database.

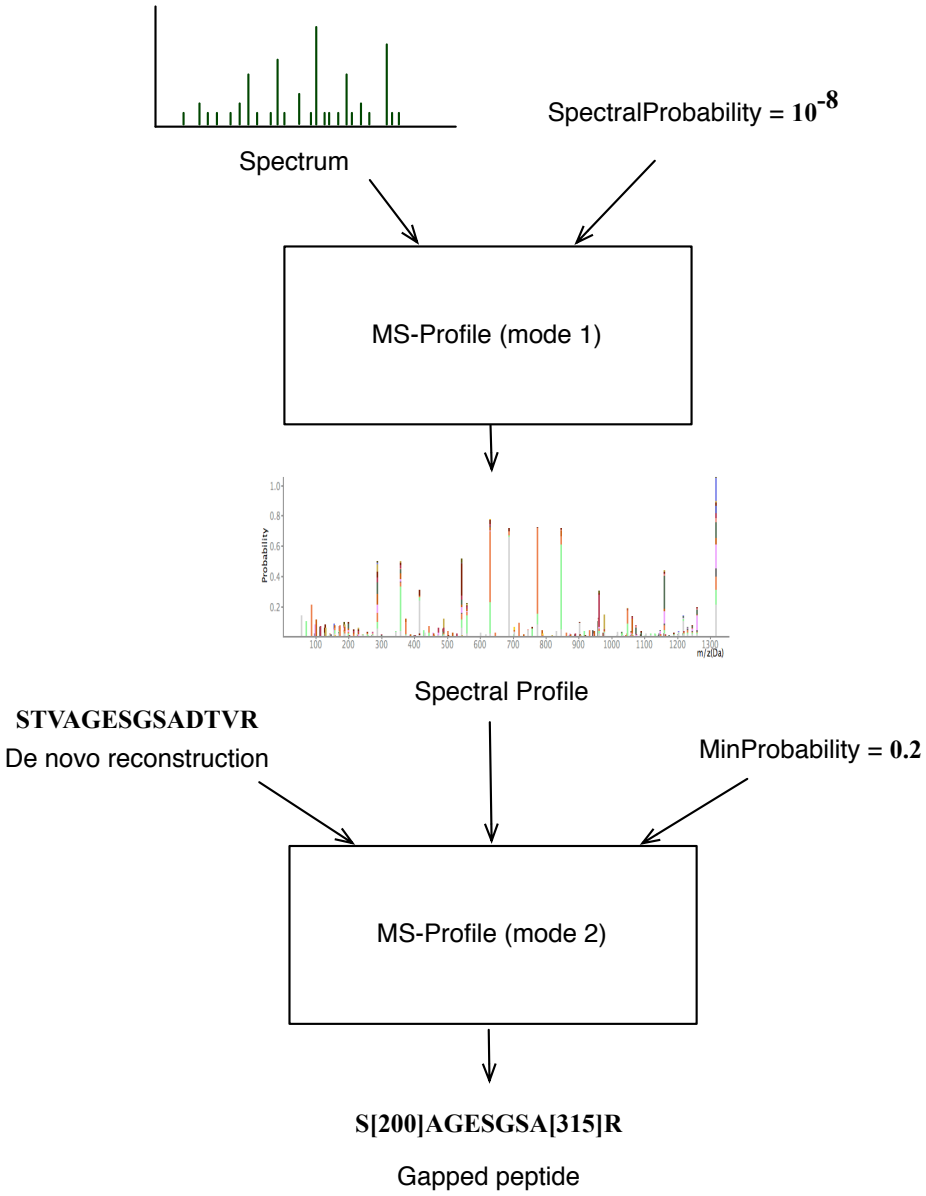
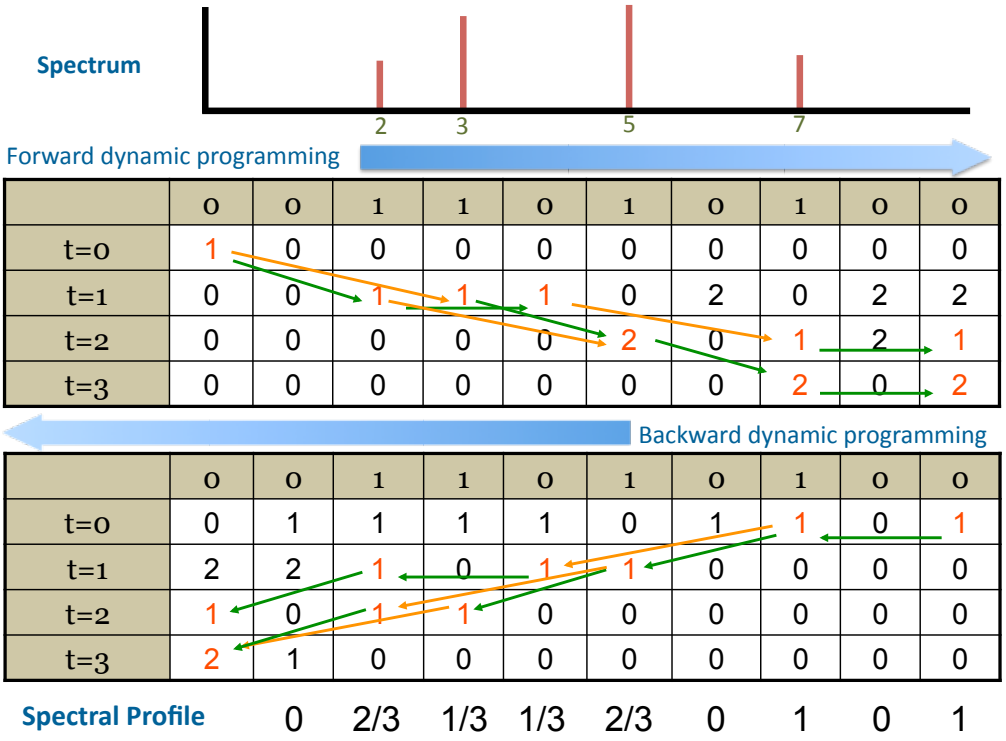


Figure 4.3: Overview of the MS-Profile tool. MS-Profile works in two modes: mode 1 is for the spectral profile generation and mode 2 is for the gapped peptide generation.



	0	0	1	1	0	1	0	1	0	0
t=0	1	0	0	0	0	0	0	0	0	0
t=1	0	0	1	1	1	0	2	0	2	2
t=2	0	0	0	0	0	2	0	1	2	1
t=3	0	0	0	0	0	0	0	2	0	2

	0	0	1	1	0	1	0	1	0	0
t=0	0	1	1	1	1	0	1	1	0	1
t=1	2	2	1	0	1	1	0	0	0	0
t=2	1	0	1	1	0	0	0	0	0	0
t=3	2	1	0	0	0	0	0	0	0	0
Spectral Profile		0	2/3	1/3	1/3	2/3	0	1	0	1

Figure 4.4: An example of the dynamic programming algorithm for computing the spectral profile of a "toy" boolean spectrum 011010100 with four peaks at masses 2, 3, 5, and 7 Da (parent mass 9). MS-Profile algorithm is illustrated with the help of a toy amino acid model (only two amino acids with masses 2 and 3 Da) and a simplified discretized spectrum. The scoring function $Score(Peptide, Spectrum)$ used for this illustration is the number of matching peaks between boolean peptide and boolean spectrum. There are only five peptides with parent mass 9: 3222 (score 3), 2322 (score 3), 2232 (score 2), 2223 and 333 (score 1). These five peptides correspond to 9-dimensional boolean vectors: 001010101, 010010101, 010100101, 010101001 and 001001001. If one considers all peptides with scores 1 and above, the spectral profile is a 9-dimensional vector $(0, \frac{3}{5}, \frac{2}{5}, \frac{2}{5}, \frac{2}{5}, \frac{2}{5}, \frac{3}{5}, 0, 1)$ representing the center of gravity of these 5 vectors. However, if one consider the dictionary of all peptides with scores 2 and above then the spectral profile $(0, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, 0, 1, 0, 1)$ is the center of gravity of 3 peptides 001010101, 010010101, 010100101. The forward-backward dynamic programming generates the spectral profile without explicitly generating any of the peptides in the dictionary. For the threshold 1 (peptides of scores 2 and above are considered), the size of the spectral dictionary is 3 and the spectral profile of the dictionary is $(0, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, 0, 1, 0, 1)$. Numbers in red and green (mass 2) and yellow (mass 3) arrows represent paths to reach the dictionary with the threshold 1.

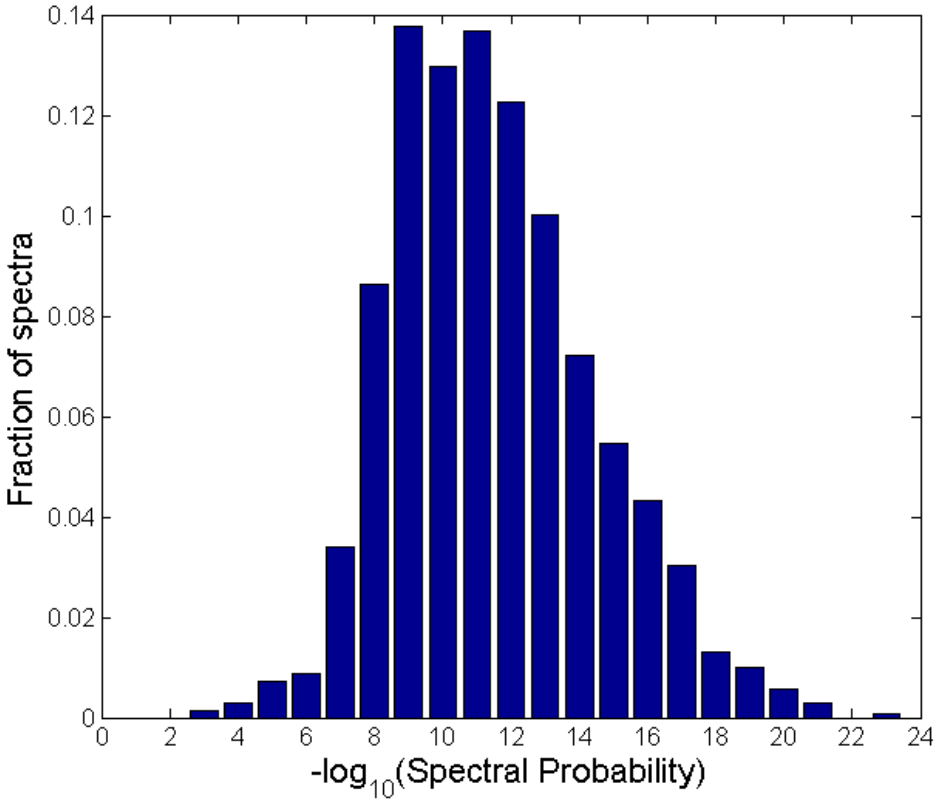
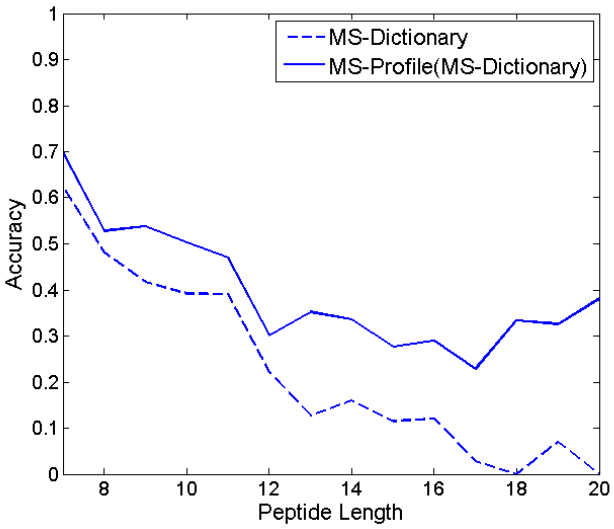
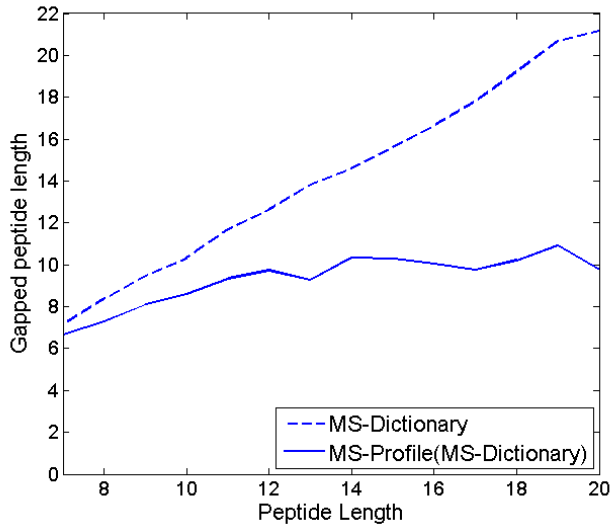


Figure 4.5: Distribution of spectral probabilities for PSM in the Standard dataset. A bar at position i represents the portion of spectra with spectral probabilities varying from $10^{-i-0.5}$ to $10^{-i+0.5}$.

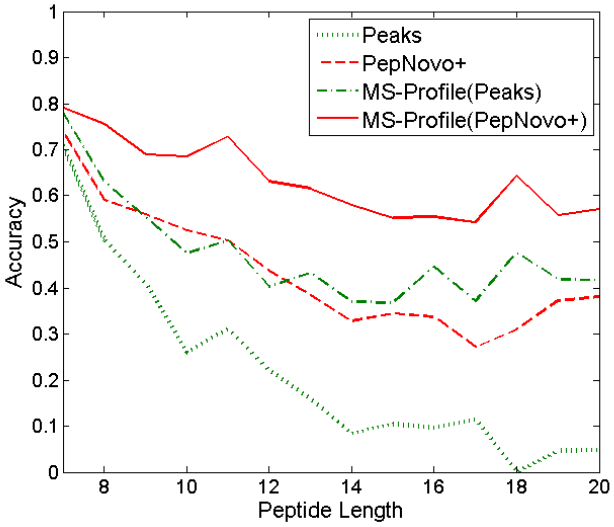


(a)

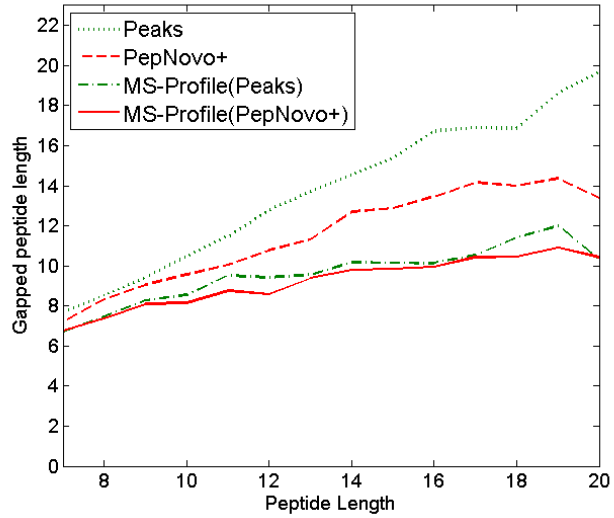


(b)

Figure 4.6: (a) Accuracy of best-scoring reconstructions generated by MS-Dictionary and the gapped peptide generated by MS-Profile (MS-Dictionary) for different peptide lengths ($MinProbability = 0.1$) If the length of a gapped peptide is less than 5 (5.6% of the spectra in the Standard dataset), we counted it as incorrect (even if the gapped peptide is correct) to penalize very short gapped reconstructions. (b) Average length of reconstructions generated by MS-Dictionary and the gapped peptides generated by MS-Profile (MS-Dictionary) for various peptide lengths.

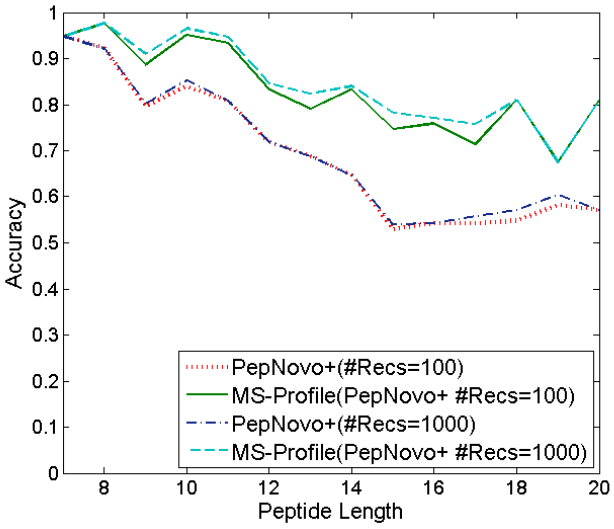


(a)

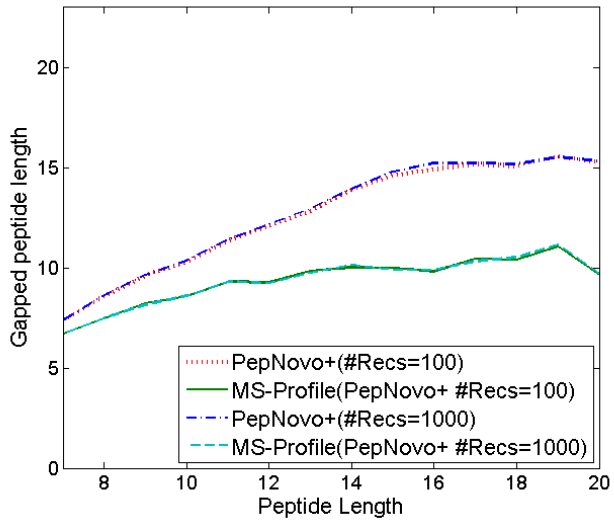


(b)

Figure 4.7: (a) Accuracy of best-scoring reconstructions generated by Peaks, PepNovo+, MS-Profile(Peaks) and MS-Profile(PepNovo+) for different peptide lengths. The reconstructions are converted into gapped peptides using MS-Profile with $MinProbability = 0.1$. If the length of a gapped peptide is less than 5, we consider it incorrect. (b) Average length of best-scoring reconstructions and gapped peptides for different peptide lengths. The length of PepNovo+ reconstructions is not proportional to the peptide length because PepNovo+ allows variable length gaps at the start and end of the peptide.

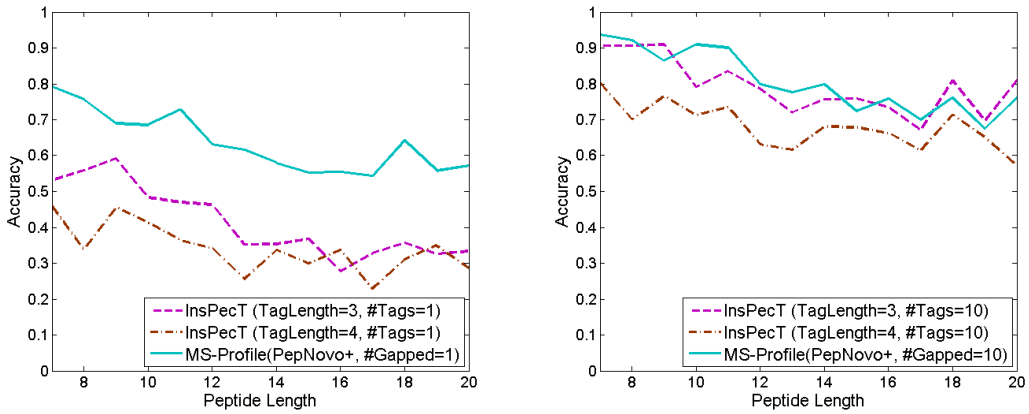


(a)



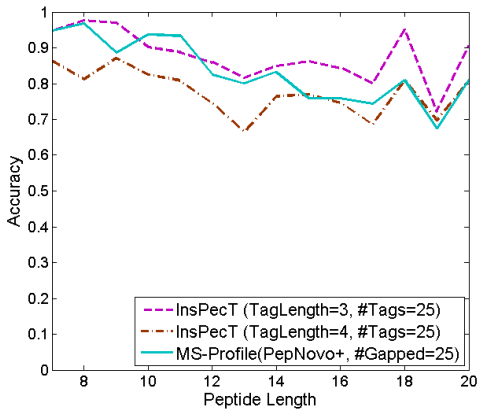
(b)

Figure 4.8: Accuracy (a) and length (b) of top 100 PepNovo+ reconstructions (PepNovo+(#Recs=100)), top 1000 PepNovo+ reconstructions (PepNovo+(#Recs=1000)), MS-Profile gapped peptides converted from top 100 PepNovo+ reconstructions (MS-Profile(PepNovo+ #Recs=100)) and MS-Profile gapped peptides converted from top 1000 PepNovo+ reconstructions (MS-Profile(PepNovo+ #Recs=1000)).



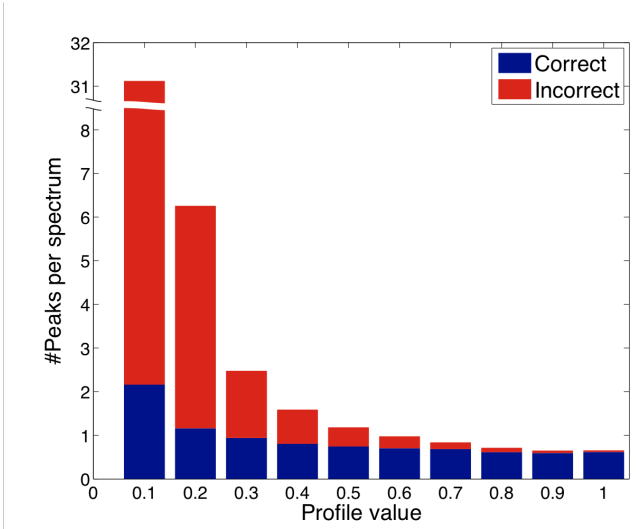
(a)

(b)

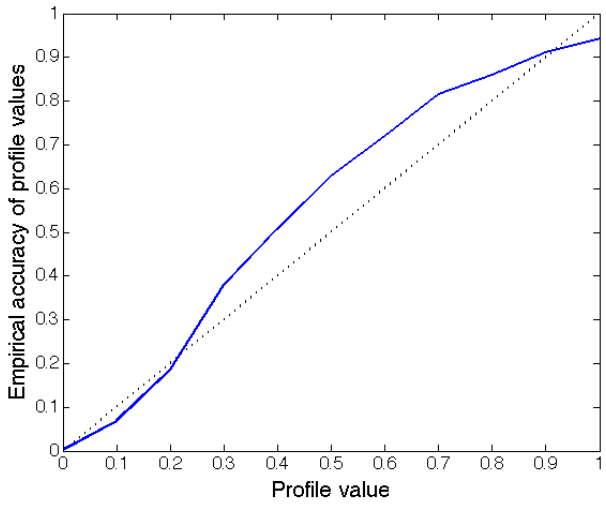


(c)

Figure 4.9: Comparison of accuracy of InsPecT tags and gapped peptides generated by MS-Profile. InsPecT (release 20080404) was run with parent mass and fragment mass tolerance 0.5Da, fixed modification of C+57, no optional modifications and without any enzyme preference. Same number of InsPecT tags of length 3, InsPecT tags of length 4 and MS-Profile(PepNovo+) gapped peptides are generated and their accuracies are shown. (a) accuracy of 1 tag or gapped peptide. (b) accuracy of 10 tags or gapped peptides. (c) accuracy of 25 tags or gapped peptides.



(a)



(b)

Figure 4.10: (a) The distribution of the average number of profile peaks per spectrum for different profile values generated by MS-Profile. The number of correct peaks is represented by blue bars; the number of incorrect peaks is represented by red bars stacked on the blue bars. A peak at position p corresponds to the profile values between $p-0.05$ and $p+0.05$. (b) The empirical accuracy (the number of correct profile peaks divided by the number of total profile peaks) of profile peaks for different profile values. The diagonal line is shown for reference.

Chapter 5

Database Search of CID, ETD, and CID/ETD Pairs

5.1 Introduction

Since the introduction of electron capture dissociation (ECD) in 1998 [102], electron-based peptide dissociation technologies have played an important role in analyzing intact proteins and post-translational modifications [103]. However, until recently, this research-grade technology was available only to a small number of laboratories since it was commercially unavailable, required experience for the operation, and could be implemented only with expensive FT-ICR instruments. The discovery of electron-transfer dissociation (ETD) [104] enabled an ECD-like technology to be implemented in (relatively cheap) ion-trap instruments. Nowadays, many researchers are employing the ETD technology for tandem mass spectra generation [105, 106, 107, 108, 109, 110].

While the hardware technologies to generate ETD spectra are maturing rapidly, software technologies to analyze ETD spectra are still in infancy. There are two major approaches to analyzing tandem mass spectra: de novo sequencing and database search. Both approaches find the best-scoring peptide either among all possible peptides (de novo sequencing) or among all peptides in a protein database (database search). While de novo sequencing is emerging as an

alternative to database search, database search remains a more accurate (and thus preferred) method of spectral interpretation, so here we focus on the database search approach.

Numerous database search engines are currently available, including SEQUEST [24], Mascot [5], OMSSA [13], X!Tandem [49], and InsPecT [25]. However, most of them are inadequate for the analysis of ETD spectra because they are optimized for collision induced dissociation (CID) spectra that show different fragmentation propensities than those of ETD spectra. Additionally, the existing MS/MS tools are biased towards the analysis of tryptic peptides because trypsin is usually used for CID, and thus not suitable for the analysis of non-tryptic peptides that are common for ETD. Therefore, even though some database search engines support the analysis of ETD spectra (e.g. SEQUEST, Mascot and OMSSA), their performance remains suboptimal when it comes to analyzing ETD spectra. Recently, an ETD-specific database search tool (Z-Core) was developed; however it does not significantly improve over OMSSA [111].

We present a new database search tool (MS-GFDB) that significantly outperforms existing database search engines in the analysis of ETD spectra, and performs equally well on non-tryptic peptides. MS-GFDB employs the generating function approach (MS-GF) that computes rigorous p-values of PSMs based on the spectrum-specific score histogram of all peptides [3].[1] MS-GF p-values are dependent only on the PSM (and not on the database), thus can be used as an alternative scoring function for the database search.

Computing p-values requires a scoring model evaluating qualities of PSMs. MS-GF adopts a probabilistic scoring model (MS-Dictionary scoring model) described in Kim et al., 2009 [3], considering multiple features including product ion types, peak intensities and mass errors. To define the parameters of this scoring model, MS-GF only needs a set of *training* PSMs.[2] This set of PSMs can be obtained in a variety of ways: for example, one can generate CID/ETD pairs and use peptides identified by CID to form PSMs for ETD. Alternatively, one can generate

[1] The term "p-value" here and the term "spectral probability" used in Kim et al., 2008 [78] are synonymous. Throughout this chapter, we use "p-value", because it is more generally used.

[2] A thousand PSMs of unique peptides is usually sufficient.

spectra from a purified protein (when PSMs can be inferred from the accurate parent mass alone) or use a previously developed (not necessary optimal) tool to generate training PSMs. From these training PSMs, MS-GF automatically derives scoring parameters without assuming any prior knowledge about the specifics of a particular peptide fragmentation method (e.g. ETD, CID, etc.) and/or proteolytic origin of the peptides. MS-GF was originally designed for the analysis of CID spectra, but now it has been extended to other types of spectra generated by various fragmentation techniques and/or various enzymes. We show that MS-GF can be successfully applied to novel types of spectra (e.g. ETD of Lys-N peptides [112, 113]) by simply re-training scoring parameters without any modification. Note that although the same scoring model is used for different types of spectra, the parameters derived to score different types of spectra are dissimilar.

We compared the performance of MS-GFDB with Mascot on a large ETD data set and found that it generated many more peptide identifications for the same false discovery rates (FDR). For example, at 1% peptide level FDR, MS-GFDB identified 9,450 unique peptides from 81,864 ETD spectra of Lys-N peptides while Mascot only identified 3,672 unique peptides, $\approx$ 160% increase in the number of peptide identifications (a similar improvement is observed for ETD spectra of tryptic peptides).[3] MS-GFDB also showed a significant 28% improvement in the number of identified peptides from CID spectra of tryptic peptides (16,203 peptides as compared to 12,658 peptides identified by Mascot).

The ETD technology complements rather than replaces CID since both technologies have some advantages; CID for smaller peptides with small charges, ETD for larger and multiply charged peptides [114, 115]. An alternative way to utilize ETD is to use it in conjunction with CID because CID and ETD generate complementary sequence information [114, 116, 117]. ETD-enabled instruments often support generating both CID and ETD spectra (CID/ETD pairs) for the same peptide. While the CID/ETD pairs promise a great improvement in peptide iden-

[3]The peptide level FDR is defined as the number of unique peptides in the decoy database over the number of unique peptides in the target database at a certain threshold. At 1% spectrum level FDR, MS-GFDB identified 22,003 spectra, while Mascot identified 9,027 spectra, a 140% increase in the number of identified spectra for ETD spectra of Lys-N peptides.

tification, the full potential of such pairs has not been fully realized yet. In the case of de novo sequencing, de novo sequencing tools utilizing CID/ETD pairs indeed result in more accurate de novo peptide sequencing than traditional CID-based algorithms [117, 118, 119]. However, in the case of database search, the argument that the use of CID/ETD pairs improves peptide identifications remains poorly substantiated. A few tools are developed to use CID/ETD (or CID/ECD) pairs for the database search but they are limited to pre-processing/post-processing of the spectral data before/after running a traditional database search tool [120, 121]. Nielsen et al., 2005 [116] pioneered the combined use of CID and ECD for the database search. Given a CID/ECD pair, they generated a combined spectrum comprised only of complementary pairs of peaks, and searched it with Mascot.[4] However, this approach is hard to generalize to less accurate CID/ETD pairs generated by ion-trap instruments since there is a higher chance that the identified complementary pairs of peaks are spurious. More importantly, using traditional MS/MS tools (like Mascot) for the database search of the combined spectrum is inappropriate, because they are not optimized for analyzing such combined spectra; a better approach would be to develop a new database search tool tailored for the combined spectrum. Recently, Molina et al., 2008 [120] studied database search of CID/ETD pairs using Spectrum Mill (Agilent Technologies, Santa Clara, CA) and came to a counterintuitive conclusion that using only CID spectra identifies 12% more unique peptides than using CID/ETD pairs. We believe that it is an acknowledgement of limitations of the traditional MS/MS database search tools for the analysis of multiple spectra generated from a single peptide.

In this chapter, we modify the generating function approach for interpreting CID/ETD pairs and further apply it to improve the database search with CID/ETD pairs. In contrast to previous approaches, our scoring is specially designed to interpret CID/ETD pairs and can be generalized to analyzing any type of multiple spectra generated from a single peptide. When CID/ETD pairs from

[4]The combined spectrum is a pseudo-spectrum generated from the set of pairs of peaks supporting the same backbone cleavage. The pair may come from the same spectrum (e.g. two peaks with the sum of masses equals to the parent mass) or different spectra (e.g. a peak from CID spectrum and a peak from ECD spectrum with the mass difference 16.02 Da, representing a possible pair of y and z fragment ions).

trypsin digests are used, MS-GFDB identified 13% and 27% more peptides compared to the case when only CID spectra and only ETD spectra are used, respectively. The difference was even more prominent when CID/ETD pairs from Lys-N digests were used, with 41% and 33% improvement over CID only and ETD only, respectively.

Assigning a p-value to a PSM greatly helped researchers to evaluate the quality of peptide identifications. We now turn to the problem of assigning a p-value to a Peptide-Spectrum-Spectrum Match (PS^2M) when two spectra in PS^2M are generated by different fragmentation technologies (e.g. ETD and CID). We argue that assigning statistical significance to a PS^2M (or even PS^nM) is a prerequisite for rigorous CID/ETD analyses. To our knowledge, MS-GFDB is the first tool to generate statistically rigorous p-values of PS^nMs.

The MS-GFDB executable is available at http://proteomics.ucsd.edu. It takes a set of spectra (CID, ETD or CID/ETD pairs) and a protein database as an input and outputs peptide matches. If the input is a set of CID/ETD pairs, it outputs the best scoring peptide matches and their p-values (1) using only CID spectra, (2) using only ETD spectra and (3) using combined spectra of CID/ETD pairs.

5.2 Methods

5.2.1 Digestion of cell lysate

HEK293 cells were grown to confluence, harvested and resuspended in lysis buffer (50 ammonium bicarbonate, 8 M urea, Complete EDTA-free protease inhibitor mix (Roche Applied Science), 5 mM potassium phosphate, 1 mM potassium fluoride and 1 mM sodium orthovanadate) and incubated for 20 min at 4 °C. An insoluble fraction was spun down at 1,000 g for 10 min at 4 °C and the protein content of the supernatant was determined using the 2DQuant Kit (GE Healthcare). Per 1 mg of lysate 45 mM dithiothreitol were used for reduction (30 min at 50 °C) and 100 mM iodoacetamide for subsequent alkylation (30 min at RT). Trypsin digests were generated by digestion of 1 mg cell lysate with 1.25 g Lys-C

for 4 h at RT followed by dilution to 2 M urea and digestion with 15 μg Trypsin for 16 h at 37 °C. Lys-N digests were made by digestion of 1 mg cell lysate with 5 μg Lys-N for 4 hours at RT, dilution to 2 M urea, and another digestion with 5 μg Lys-N for 16 h at 37 °C.

5.2.2 Peptide pre-fractionation by strong cation exchange (SCX)

Fractionation of peptides was performed as described earlier [122, 123]. In detail, digests were acidified with formic acid and loaded onto two C18 cartridges using an Agilent 1100 HPLC system operated at 100 μl/min with 0.05% formic acid in water. Peptides were then eluted from the C18 cartridges using 80% acetonitrile and 0.05% formic acid in water onto a PolySULFOETHYL A column (200 mm x 2.1 mm column, PolyLC). Separation of different peptide populations was performed at 200 μL/min using a non-linear gradient as follows: 0 to 10 min 100% solvent A (5 mM KH_2PO_4, 30% acetonitrile, 0.05% formic acid), 10 to 15 min from 0 to 26% solvent B (350 mM KCl, 5 mM KH_2PO_4, 30% acetonitrile, 0.05% formic acid), 15 to 40 min from 26 to 35% solvent B and from 40 to 45 min from 35 to 60% solvent B, and from 45 to 49 min from 60 to 100% solvent B. Fractions were collected in 1 min intervals for 40 min, dried down in a vacuum centrifuge, and resuspended in 10% formic acid.

5.2.3 Mass spectrometry

SCX fractions were analyzed on a reversed phase nano-LC-coupled LTQ Orbitrap XL ETD (Thermo Fisher Scientific). An Agilent 1200 series HPLC system was equipped with a 20 mm Aqua C18 (Phenomenex) trapping column (packed in-house, 100 μm inner diameter, 5 μm particle size) and a 400 mm ReproSil-Pur C18-AQ (Dr. Maisch GmbH) analytical column (packed in-house, 50 μm inner diameter, 3 μm particle size). Trapping was performed at 5 μL/min solvent C (0.1 M acetic acid in water) for 10 min, and elution was achieved with a gradient from 10 to 30% (v/v) solvent D (0.1 M acetic acid in 1:4 acetonitrile : water) in solvent

C in 110 min, followed by a gradient of 30 to 50% (v/v) solvent D in solvent C in 30 min, followed by a gradient of 50 to 100% (v/v) solvent D in solvent C in 5 min and finally 100% solvent D for 2 min. The flow rate was passively split from 0.45 mL/min to 100 nL/min. Nano-electrospray was achieved using a distally coated fused silica emitter (360 μm outer diameter, 20 μm inner diameter, 10 μm tip inner diameter, New Objective) biased to 1.7 kV. The instrument was operated in data dependent mode to automatically switch between MS and MS/MS. Survey full scan MS spectra were acquired from m/z 350 to m/z 1500 in the Orbitrap with a resolution of 60,000 at m/z 400 after accumulation to a target value of 500,000 in the linear ion trap. The two most intense ions at a threshold of above 500 were fragmented in the linear ion trap using CID at an AGC target value of 30,000 and ETD with supplemental activation at an AGC target value of 50,000. The ETD reagent AGC target value was set to 100,000 and the reaction time to 50 ms.

5.2.4 Data processing

From every raw data file recorded by the mass spectrometer, representing a single SCX fraction, two different peaklists containing either CID or ETD fragmentation data were generated using Proteome Discoverer (version 1.0, Thermo Fisher Scientific) with a signal-to-noise threshold of 3 and the following settings for the ETD-non-fragment filter: precursor peak removal with 4 Da, charge-reduced precursor removal with 8 Da, and removal of known neutral losses from charge-reduced precursors with 8 Da within a window of 120 Da. Single-fraction peaklists of the major peptide-containing SCX fractions for Trypsin-derived and Lys-N-derived peptides were then merged into four larger peaklists, denoted CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN. The whole data set is composed of 168,960 CID/ETD pairs. 87,096 pairs (51,233 with charge 2+, 24,854 with charge 3+ and 11,009 with charges 4+ and larger) are from the Trypsin digests and 81,864 (24,284 with charge 2+, 28,168 with charge 3+ and 29,412 with charges 4+ and larger) are from the Lys-N digests. All the data sets are available on Tranche (http://proteomecommons.org/tranche/). Spectra with precursor charges from 2+ to 7+ were considered in the further analyses.

5.2.5 Mascot analysis

Mascot (version 2.3.0, Matrix Science) was used to search the peaklists against an in-house built database (74,190 entries; 31,263,418 amino acids) assembled from the IPI human database (version 3.52, http://www.ebi.ac.uk/ipi) plus common contaminants (target database). A decoy database was constructed by reversing all sequences and slightly scrambling entries using MaxQuant (version 1.0.13.8; http://www.maxquant.org) [124]. The target and decoy databases were searched separately to estimate FDRs. The following parameters were used for database searching: 50 ppm precursor mass tolerance, 0.5 Da fragment ion tolerance, up to 2 missed cleavages allowed, carbamidomethyl cysteine as fixed modification, no variable modifications. The enzyme was specified as either Trypsin or Lys-N and the instrument type either ESI-TRAP or ETD-TRAP.

5.2.6 MS-GF training

MS-GF takes a set of peptide-spectrum matches (PSMs) as a training set and outputs a file containing scoring parameters. All spectra in the training set are assumed to be generated using the same fragmentation method and the same enzyme. Below we describe the following 5 steps for generating MS-GF scoring parameters: (1) partitioning the training set, (2) selecting precursor offsets for removal, (3) selecting ion types, (4) computing peak rank distributions and (5) computing peak error distributions. We remark that steps (2) and (3) were missing in the previous MS-GF version [78] thus forcing users to specify the ion types manually. Note that by adding the steps (2) and (3), MS-GF can now automatically learn scoring parameters from any type of spectra with any precursor charges.

Partitioning the training set

Fragmentation propensities of mass spectra strongly depend on the precursor charge [125] and the peptide length [3].[5] Therefore, we use different sets of

[5]The differences in the fragmentation propensity between peptides of similar lengths (like 7 and 8 amino acids) are typically small as compared with differences between peptides with very different lengths (like 7 and 20 amino acids).

scoring parameters depending on the precursor charge, and the peptide length. To generate the scoring parameters, we partition the training set by the precursor charge of the spectrum and the estimated peptide length inferred from the precursor mass of the spectrum. Then, for each partition, we learn the parameters using only spectra belonging to this partition. In addition, since different types of peaks have different propensities with respect to the relative positions in the spectrum (e.g. peaks corresponding to doubly charged ions only appear in the lower part of the spectrum), we learn the parameters separately for the lower and the upper halves of the mass range.

Selecting precursor offsets for removal

ETD spectra often possess precursor peaks, charge-reduced precursor peaks, their neutral losses and side-chain losses [126]. While those peaks usually have high intensities, they do not contribute useful information for peptide identifications. We therefore remove these peaks to avoid a risk of erroneously interpreting them as other ion types. To figure out which the peaks have to be removed, we use the offset frequency function (OFF) [33]. OFF is a histogram of the peaks observed at a relative offset from a specific m/z in the spectrum. Here we use the OFFs from the precursor mass and charge-reduced precursor m/z's.

First we filter all spectra in the training set to remove noisy peaks as follows: given a peak at mass m, we retain the peak if it is among the top k (k = 6 by default) peaks within a window of size 100 Da around m. Then we compute the OFFs from the precursor m/z and all possible charge-reduced precursor m/z's. If a certain offset is observed in more than a predefined portion of the spectra (15% by default), we mark the offset for removal. Later, all peaks observed at marked offsets are filtered out (see Figure 5.3).

Selecting ion types

For each partition of the training set, we select ion types to be used for scoring using the OFF of the prefix and suffix residue masses as described in Dancik et al., 1999 [33]. We represent an ion type by a triplet of (charge, prefix or

suffix, offset) and consider all possible prefix and suffix ions with charges 1 to the precursor charge and integer offsets from -38 to +38. If we observe an ion type at more than a predefined portion of all cleavage sites in the filtered spectra (15% by default), we select the ion type.

Computing peak rank distributions

For each selected ion type at a certain partition, we compute the probability of a peak of rank i being the ion type (ion rank probability) from rank 1 to $MaxRank$ (150 by default). We also compute the probability of a peak of rank i being an ion type that is not selected (noise rank probability). As described in [3], the log of the ion rank probability over the noise rank probability at certain rank serves as the *rank score* of a peak.

Computing peak error distributions

Instead of setting up a fixed mass error threshold (e.g. 0.5 Da for ion traps) and assigning the same score to all peaks within this error threshold, we vary scores depending on the mass error. To do this, for each selected ion type at each partition, we compute the mass error histogram of all peaks of ranks within $MaxRank$ assigned to that ion type (ion error probability). We also compute a similar histogram using ion types that are not selected (noise error probability). The log ratio of the ion error probability over the noise error probability serves as the *error score* of a peak.

Training scoring parameters for CID-Tryp, ETD-Tryp, CID-LysN, and ETD-LysN

We first generated initial scoring parameter files for the four data sets (CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN) using PSMs with Mascot scores corresponding to peptide level FDRs less than 1% as a training set. Using these initial parameter files, we ran MS-GFDB and selected PSMs with MS-GF p-values corresponding to peptide level FDRs less than 1%. These PSMs were used as a new training set to build the final scoring parameter files.

5.2.7 MS-GFDB search (for CID or ETD spectra)

Since MS-GFDB automatically pre-processes spectra, we converted each raw data file into an mzXML file using ReAdW 4.3.1 [127] and used the mzXML file in the MS-GFDB search (as opposed to using Proteome Discoverer for noise and (charge-reduced) precursor filtering). MS-GFDB searches were carried out against the same database with the same parameters as were used for Mascot searches.

MS-GFDB uses two scores: the MS-GF score and the p-value (both are computed by MS-GF). The MS-GF score is used to evaluate the quality of a PSM and the p-value is used to assess the statistical significance of a PSM. To compute the MS-GF score, MS-GF first converts every spectrum into a Prefix-Residue Mass (PRM) spectrum [25, 33] using scoring parameters specific to a particular fragmentation technique and enzyme. The PRM spectrum is a scored version of a spectrum having a score at every mass up to the parent mass of the spectrum.[6] As described in Dančík et al., 1999 [33], the score of a PRM spectrum at mass m represents the log likelihood ratio that the peptide from which the spectrum was derived contains a prefix of mass m.[7] The MS-GF score of a peptide against a spectrum is defined as the sum of scores in the PRM spectrum corresponding to prefix masses of the peptide. To compute the p-value, MS-GF generates the score histogram of *all peptides* using the generating function approach (see [78] for details on the generating function approach). The p-value of a peptide with match score s is defined as the area under the histogram where the score value (x-axis) is equal or larger than s. Fig. 5.1 illustrates the procedure to compute p-values with MS-GF.

Given a spectrum and a protein database, MS-GFDB computes MS-GF scores for all the peptides in the database (similarly to SEQUEST or Mascot), finds the peptide with the best score and reports its p-value.[8]

[6] One can define the granularity of a mass depending on the resolution of the mass spectrum. Throughout this chapter, the granularity is set as 1 Da (equivalent to the fragment ion tolerance 0.5 Da). While this chapter focuses on MS/MS spectra with inaccurate fragment masses, MS-GFDB can be adapted to analyze spectra with accurate fragment masses by changing the granularity.

[7] Every peptide of length n defines $n-1$ *prefix* masses representing masses of the first i amino acids (for $1 \leq i < n$).

[8] MS-GFDB search takes only ≈ 0.1 second per spectrum against a database containing 31 mil-

5.2.8 MS-GFDB search (for CID/ETD pairs)

MS-GFDB combines a pair of tandem mass spectra generated from a single precursor ion (using different fragmentation techniques) and matches the combined spectrum against a database. Given a pair of spectra, it first converts each spectrum into a PRM spectrum (using fragmentation-specific parameters for each type of spectrum) and generates a *summed PRM* spectrum. The *Summed PRM spectrum* of two PRM spectra (with the same parent mass) is calculated by *adding* two PRM scores (log likelihood ratios) corresponding to the same mass. For example, if at mass 500, two PRM spectra have scores 7 and 3, correspondingly, the summed PRM spectrum has score 7+3=10 at mass 500. Note that summing PRM scores at mass m is equivalent to multiplying the probabilities that mass m is a prefix mass of the peptide from which each spectrum was derived. This summed PRM model assumes that ion types are independent within the same spectrum [129] and when coming from different spectra [79], the assumption that proved to be useful in other applications. The score histogram of a CID/ETD pair is computed using the summed PRM spectrum and is used to compute p-values. Fig. 5.2 illustrates the flow of the p-value computation for CID/ETD pairs. This method improves on the previous method proposed by Nielsen et al. [116] in that it merges evidence for a certain backbone cleavage (represented as a PRM score) using a probabilistic model, whereas the approach in [116] only retains a peak if it has a complementary peak or discards a peak if not. Therefore, the approach in [116] results in much stricter peak filtering, making it difficult to distinguish between correct and incorrect peptide identifications. For example, given a CID/ETD pair with a poor-quality CID spectrum and a high-quality ETD spectrum, the method in [116] is unlikely to interpret the pair, since the CID spectrum does not help to identify "complementary pairs of peaks" and the resulting spectrum contains only a few peaks identified from the ETD spectrum itself. In contrast, the summed PRM

lion amino acids for a computer with Core i7 2.7Ghz CPU with 12GB memory. We have recently published a study to further speed up MS-GFDB using gapped peptides (MS-GappedDictionary, Jeong et al., 2010 [128]), an approach that is similar to using peptide sequence tags in Inspect [25]. MS-GappedDictionary uses MS-GF scores to generate gapped peptides that are used for fast database scan like peptide sequence tags. Combining MS-GappedDictionary and MS-GFDB enables orders of magnitudes speed-up.

scores retain most of the sequence information in the ETD spectrum contributing to successful peptide identifications.

Note that this method can be generalized to the case of analyzing more than two tandem mass spectra generated from a single precursor ion (e.g. by adding a high energy collisional dissociation (HCD)/beam-type CID spectrum).

5.3 Results

5.3.1 Analysis of individual spectra

For each of the CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN data sets, we compared the performance of MS-GFDB with Mascot by counting the number of identified peptides for each FDR (peptide-level FDR) using the separate target-decoy search approach [130]. For all the four data sets, MS-GFDB outperformed Mascot (Fig. 5.4). For example, at 1% FDR, MS-GFDB identified 14,409 peptides in ETD-Tryp while Mascot identified 5,310 peptides. The difference is more notable for ETD spectra than CID spectra and for Lys-N digests than trypsin digests. This indicates that Mascot is poorly optimized for the analysis of new data types while MS-GFDB automatically adapts to novel types of data. Even in the case of the CID-Tryp data set where Mascot has been subjected to a decade-long development, MS-GFDB identified $\approx$ 30% more peptides across entire FDR range. Similar results were obtained using the spectrum-level FDR.

MS-GFDB also outperformed SEQUEST and OMSSA (data not shown). To boost the performance of existing MS/MS database search tools, PeptideProphet [11], iProphet and Percolator [131, 132] rescore their PSMs, resulting in a significant increase in the number of peptide identifications [133]. However, MS-GFDB outperformed even PeptideProphet, iProphet and Percolator which take advantage of extra information unavailable to MS-GF such as the score distribution of all PSMs and the retention time information (Figure 5.5 and Figure 5.6).

In this experiment, we used the same data for both training and testing of the performance, thus raising a valid concern about over-fitting. This was done because we observed that MS-GF parameters characterize a particular protocol

(e.g. ETD for a particular enzyme) and are rather stable with respect to specific data sets, i.e. variable data sets with the same protocol result in similar MS-GF parameters. To address this concern, we demonstrated that if we derive MS-GF scoring parameters from a training data set A and apply it to a test data set B, the results hardly change as compared to deriving MS-GF scoring parameters from the data set B and apply it to the same data set B (see Figure 5.7).

For further analyses below, PSMs with FDRs below 1% were selected from the four data sets using MS-GFDB; if multiple spectra of the same charge are matched to the same peptide, only that with the best score was chosen. From CID-Tryp/ETD-Tryp/CID-LysN/ETD-LysN data set, 16,203/14,409/8,893/9,450 PSMs were selected and denoted by CID-Tryp-Confident/ETD-Tryp-Confident/ CID-LysN-Confident/ETD-LysN-Confident.

5.3.2 Comparison of ion fragmentation statistics across different spectral data sets

The spectra of the same peptide are different depending on the fragmentation methods and precursor ion charges. Moreover, spectra of peptides produced by one enzyme (e.g. tryptic peptides ending with Lys or Arg) do have different fragmentation propensities than spectra of peptides produced by other enzyme (e.g. Lys-N peptides starting from Lys) [122, 134]. The common knowledge that ETD spectra are mainly comprised of c and z· ions (and their neutral losses) while CID spectra are of b and y ions (and their neutral losses) is insufficient for designing a good scoring function since one has to know the propensities (likelihood) of these ions and many other neutral losses [135]. To analyze such propensities for different types of spectra, we measured the probability of a certain ion type being observed (Fig. 5.8) and plotted the distribution of a peak of a given rank being a certain ion type (Fig. 5.9 and 5.10) as presented in [3, 33].[9] Note the high abundance of c ions with high intensities in Fig. 5.9 (d), confirming the previously published result [112]. Features shown in Fig. 5.8, 5.9 and 5.10 were automatically

[9]Rank of a peak is defined as the number of peaks (in the same spectrum) with intensities higher than or equal to intensity of the peak [3].

derived by MS-GF scoring functions and contributed to the improved performance of MS-GFDB over other tools.

5.3.3 Pitfalls of "intersection" and "union" approaches to identifying CID/ETD pairs

It is believed that utilizing CID/ETD pairs is helpful to improve confidence of peptide identifications since the identification from one method cross-validates the other. However, there is no consensus on how to utilize CID/ETD pairs for the database search. The common practice is to run database search for CID spectra and ETD spectra separately as if the pairing is not even known, identify confident PSMs using a pre-defined threshold (e.g. peptide level FDR 1% or a pre-defined score threshold) and take the intersection of CID PSMs and ETD PSMs (intersection approach). For example, in CID-Tryp and ETD-Tryp there are 50,765 spectral pairs where either CID or ETD spectra (or both) are confidently identified with MS-GFDB within the peptide level FDR 1%. In 32,431 spectral pairs (representing 12,093 distinct peptides), the CID identification and ETD identification were the same, indicating that these identifications are reliable (Fig. 5.11 (a)). To measure the FDR of these "intersection" spectral pairs, we repeated the same procedure with the identifications to the decoy database and obtained 8 pairs (representing 5 peptides) where CID and ETD identifications agree (Fig. 5.11 (b)); hence, the peptide level FDR corresponds to $5/12,093 = 4.1 \cdot 10^{-4}$. While taking the intersection improved the confidence of the resulting peptide identifications (12,093 peptides at FDR close to 0), at the same confidence level, MS-GFDB identified 7% more peptides using only CID spectra (not shown in Fig. 5.11)![10] This indicates that this approach is inefficient considering the half of the instrument time was wasted generating ETD spectra that did not help to improve the number of peptide identifications.

The poor performance of the intersection approach can be explained by the

[10]For Lys-N digests, we identified 5,788 peptides using the intersection approach with a corresponding to $3.5 \cdot 10^{-4}$ FDR; at the same FDR, MS-GFDB identified a similar number of peptides using only CID spectra.

dependencies in scores of CID and ETD spectra from the same pair. Examination of hits in the decoy database revealed that a high scoring PSM for CID spectra often corresponds to a high scoring PSM for ETD spectra from the same pair. As a result, contrary to the common belief, the intersection approach has limited ability to remove incorrect PSMs. On the other hand, many hits in the target database have high scores for CID spectra and low scores for ETD spectra (or vice versa), thus reducing the number of correct PSMs returned by the intersection approach.

Similarly, it is possible to take the "union" of identified peptides (all significant CID identifications plus all significant ETD identifications) to get more peptide identifications. For instance, from the above 50,765 spectral pairs, one may take the 4,073+12,093+2,280 = 18,446 peptides, corresponding to FDR (154 +5+137)/18,446 = 1.6%.[11] At the same FDR level, MS-GFDB identified 16,636 peptides only from CID spectra, thus this union approach resulted in 11% increase in the number of peptides. While this improvement in the number of peptides (with a larger FDR) is meaningful, our proposed approach results in a comparable number of identified peptides at a stricter level of confidence (1% FDR instead of 1.6%).

5.3.4 Identifications from combined CID/ETD spectra

Given a CID/ETD pair, one can generate a "combined spectrum" and search a database with the combined spectrum. We used the summed PRM spectra as described above (denoted by MS-GFDB CID/ETD) and compared its performance with MS-GFDB using only CID spectra (MS-GFDB CID) or ETD spectra (MS-GFDB ETD). MS-GFDB CID/ETD identified more peptides across entire FDR range compared to MS-GFDB CID or MS-GFDB ETD for both trypsin digests and Lys-N digests (Fig. 5.12). For example, at 1% FDR, MS-GFDB CID/ETD identified 18,342 peptides from CID/ETD pairs of trypsin digests and 12,561 peptides from LysN digests, corresponding to 13%, 27%, 41% and 33% improvement over when CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN data sets

[11] Spectral pairs where CID and ETD identifications disagree (red numbers in Fig. 5.11) were discarded.

are separately used, respectively. If we consider spectra of charge 3 and larger (where ETD has advantages over CID), the improvement becomes even more significant: 23%, 30%, 68% and 21%.

The improved performance of MS-GFDB CID/ETD is due to the probabilistic model for constructing combined spectra. We remark that a brute-force approach to constructing combined spectra actually reduces the number of peptide identifications.

5.4 Discussion

We demonstrated that the generating function approach is easily adaptable to the analysis of novel types of spectra. For all types of spectral data sets we have tested, MS-GFDB outperformed state-of-the-art MS/MS database search tools. We further demonstrated how to utilize the combined CID/ETD spectra generated from CID/ETD pairs using MS-GFDB.

We emphasize that MS-GFDB analyzes all different data sets in exactly the same way using different scoring parameters that are automatically derived by the same training procedure. While it may seem counterintuitive that the MS-GF scoring function (defined as a simple dot-product of vectors) improves on more complex scoring functions used in traditional MS/MS tools, it was made possible by deriving rigorous MS-GF p-values using the generating function approach. We are not claiming that MS-GF scores are "better" than Mascot scores, but we do show that p-values derived from MS-GF scores greatly improve on Mascot scores. This observation emphasizes the importance of rigorous p-values that remain unavailable for popular tools like Mascot and SEQUEST.

The problem of analyzing spectral pairs from the same precursor is related to the problem of combining database search scores of MS^2 and MS^3 spectra from the same peptide addressed by Olsen and Mann, 2004 [136], Bandeira et al., 2008 [79] and Ulintz et al., 2008 [137]. Olsen and Mann, 2004 and Bandeira et al., 2008 developed a probabilistic scoring model for MS^3 spectra and used it to adjust the MS^2 score by summing the MS^2 and MS^3 scores. While this approach is

similar to our approach in that both use the sum of (log-likelihood) scores as the score of a pair, it did not provide a rigorous framework to compute the p-value of the pair. On the other hand, Ulintz et al., 2008 developed an approach searching the database separately for MS^2 and MS^3 spectra and adjusting the probabilities of both spectra if the top scoring sequences match (similar to the intersection approach described above). In contrast, our approach considers all possible cases (e.g. including peptides having poor scores against CID spectrum and good scores against ETD spectrum) and uses them to compute p-values, something that was missing in previous studies.

ETD has certain advantages over CID in the analysis of peptides with post-translational modifications (PTMs) [112, 138, 139, 140]. MS-GFDB can be used to identify modified peptides. When PTMs are selected in advance (restrictive search for PTMs), MS-GFDB only needs to add the masses of amino acids with PTMs to the standard 20 amino acid set. In the analysis of a sample of phosphorylated peptides, MS-GFDB identified about 30-40% more peptides from CID spectra and about 60-90% more peptides from ETD spectra than Mascot . The gain from MS-GFDB over Mascot in this data set was smaller than in the other data sets described above. This is because we used the parameters trained from unmodified spectra to score spectra of phosphorylated peptides. It is well known that some post-translational modifications (PTMs) like phosphorylation change the fragmentation propensity of the spectrum, especially in the case of CID spectra [141]. Therefore, to efficiently analyze such PTMs, one needs to develop a scoring function that is specific to the target PTM [89]. Designing a PTM-specific scoring function and the generating function for modified peptides is beyond the scope of this chapter.

5.5 Acknowledgements

This chapter, in full, was published as “The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search”. S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mo-

hammed, A. J. R. Heck, and P. Pevzner. *Molecular & Cellular Proteomics*, vol. 9, no. 12, pp. 2840-2852, 2010. The dissertation author was the primary author of this paper.

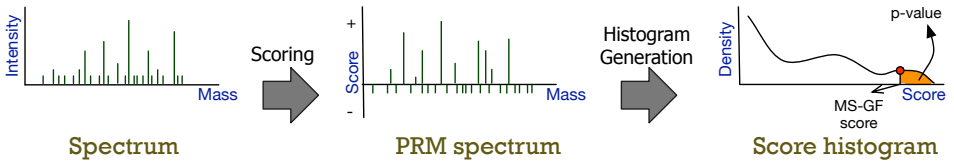


Figure 5.1: Computing p-values with MS-GF for a single spectrum. Given a tandem mass spectrum, MS-GF converts the spectrum into a PRM spectrum (scored version of the tandem mass spectrum). The score of a PRM spectrum at mass m represents the log likelihood ratio that the peptide from which the spectrum was derived contains a prefix of mass m. Negative peaks in the PRM spectrum represent masses more likely to represent incorrect rather than correct prefix masses. Such negative peaks in the PRM spectrum usually correspond to low-intensity or missing peaks in the experimental spectrum. The PRM spectrum is used to compute the MS-GF score of any peptide against the spectrum. Then, MS-GF computes the histogram of the MS-GF scores of all peptides against the spectrum using the generating function approach. Finally, MS-GF computes the p-value of a peptide as the area under the histogram with MS-GF scores equal or larger than the MS-GF score of the peptide.

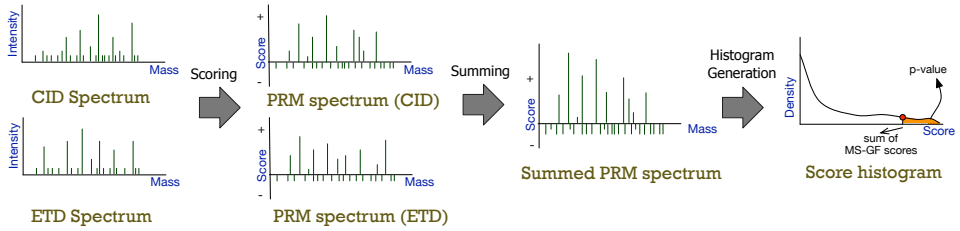


Figure 5.2: Computing p-values with MS-GF for CID/ETD pairs. Given a CID/ETD pair, MS-GFDB converts each spectrum into a PRM spectrum and merges two PRM spectra by summing scores of peaks sharing the same mass. This "summed" PRM spectrum is used to generate the score histogram of all peptides and p-values are computed using the histogram.

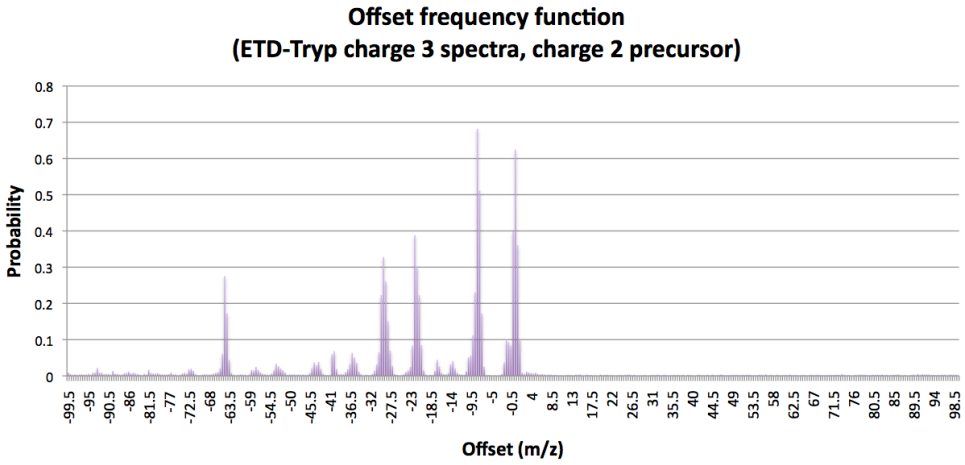


Figure 5.3: Example of the offset frequency function (OFF) from the (charge-reduced) precursor m/z. Shown is the OFF from the charge 2 (charge-reduced) precursor m/z of charge 3 spectra in ETD-Tryp data set. The horizontal axis represents the distance (in m/z) from the charge-reduced precursor m/z. The vertical axis represents the probability that a peak of the corresponding offset exists. For example, in about 40% of the charge spectra in ETD-Tryp data set, there exists a peak with m/z corresponding to the charge 2 precursor m/z minus 22. All the offsets over a predefined probability (0.15 by default) are marked for removal.

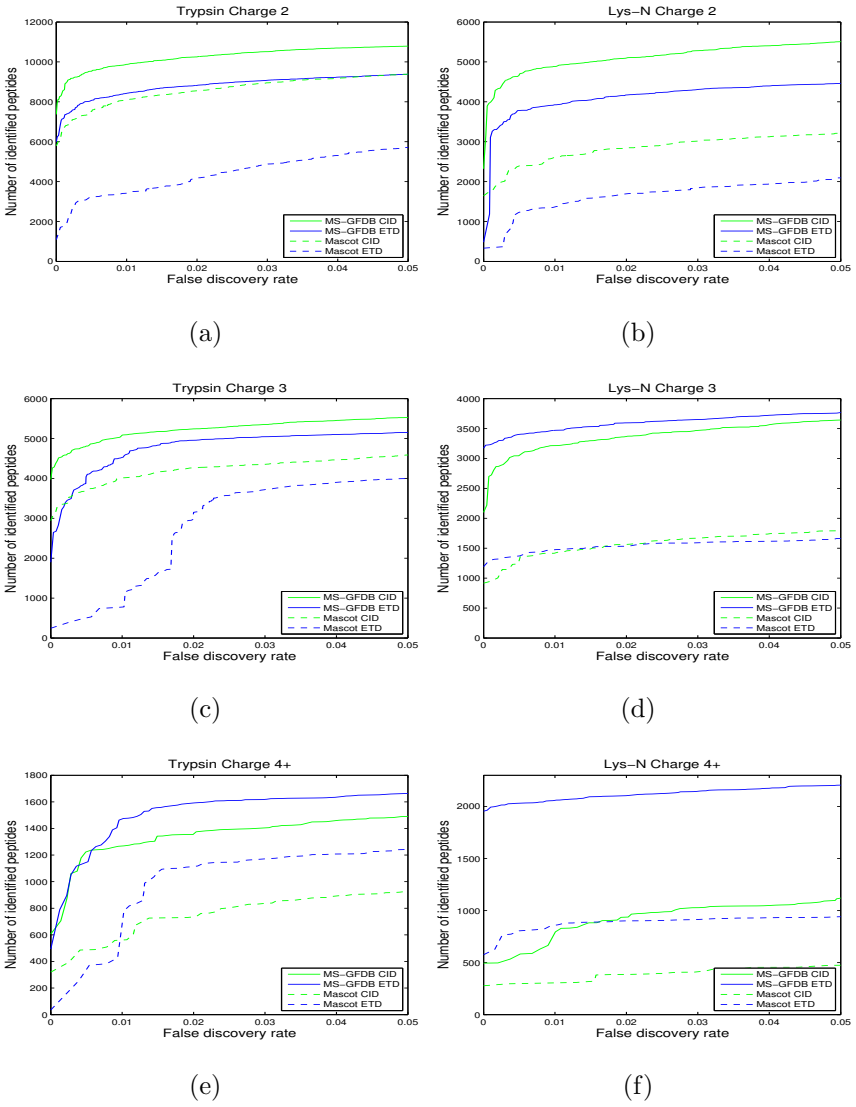


Figure 5.4: Number of identified peptides with Mascot and MS-GFDB from (a) charge 2 spectra in CID-Tryp and ETD-Tryp, (b) charge 2 spectra in CID-LysN and ETD-LysN, (c) charge 3 spectra in CID-Tryp and ETD-Tryp, (d) charge 3 spectra in CID-LysN and ETD-LysN, (e) spectra with charges 4 and larger in CID-Tryp and ETD-Tryp, and (f) spectra with charges 4 and larger in CID-LysN and ETD-LysN. The number of peptide identifications is plotted against the corresponding peptide level FDR. Solid curves represent MS-GFDB and dashed curves represent Mascot. Green curves represent CID and blue curves represent ETD. Mascot ion scores and MS-GFDB p-values were used for computing FDRs. FDRs were separately computed for spectra of precursor charge 2, precursor charge 3, and precursor charge 4 and larger. For all the cases considered, MS-GFDB outperformed Mascot.

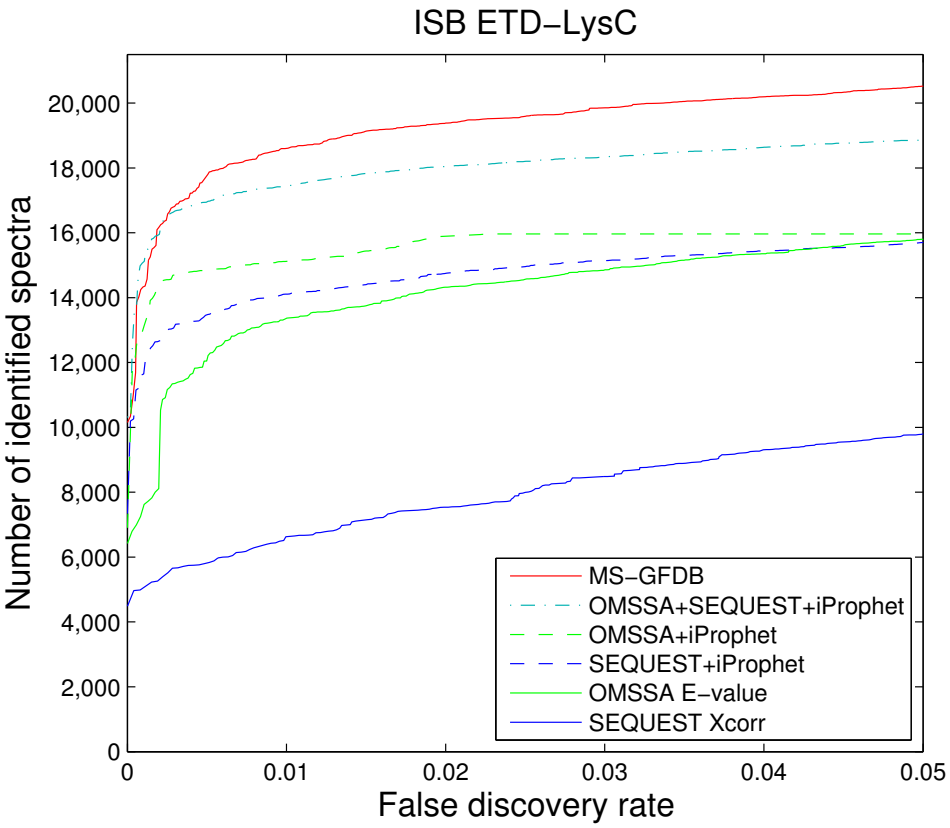
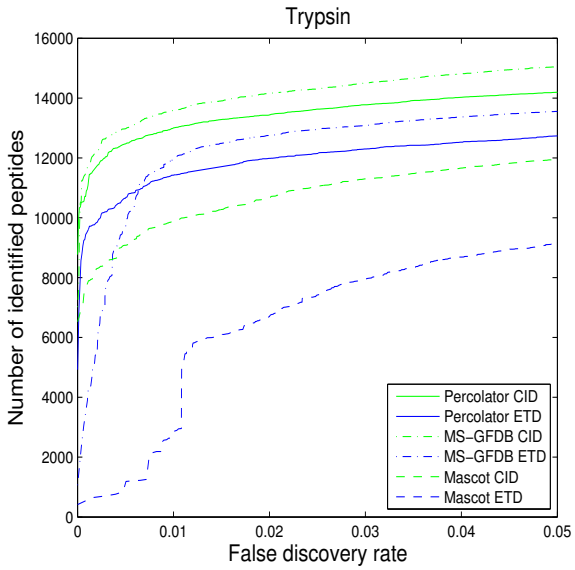
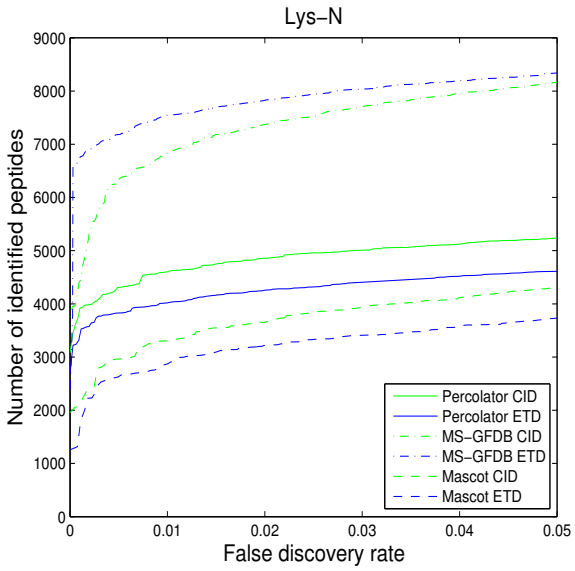


Figure 5.5: Number of identified spectra (out of 61,020 spectra) with SEQUEST, OMSSA, OMSSA+iProphet, SEQUEST+iProphet, OMSSA+SEQUEST+iProphet and MS-GFDB. False discovery rates were calculated using the combined TDA. OMSSA+PeptideProphet and SEQUEST+PeptideProphet show similar results as OMSSA+iProphet and SEQUEST+iProphet, respectively and thus are not shown. MS-GFDB outperformed all other tools.



(a)



(b)

Figure 5.6: Number of identified peptides with Percolator at varying FDRs from the (a) trypsin digests (b) Lys-N digests. MS-GFDB and Mascot results were also shown for reference. Posterior error probabilities were used to compute FDRs. Percolator significantly improved on Mascot (especially for ETD spectra from trypsin digests), but identified less peptides than MS-GFDB at the same FDR.

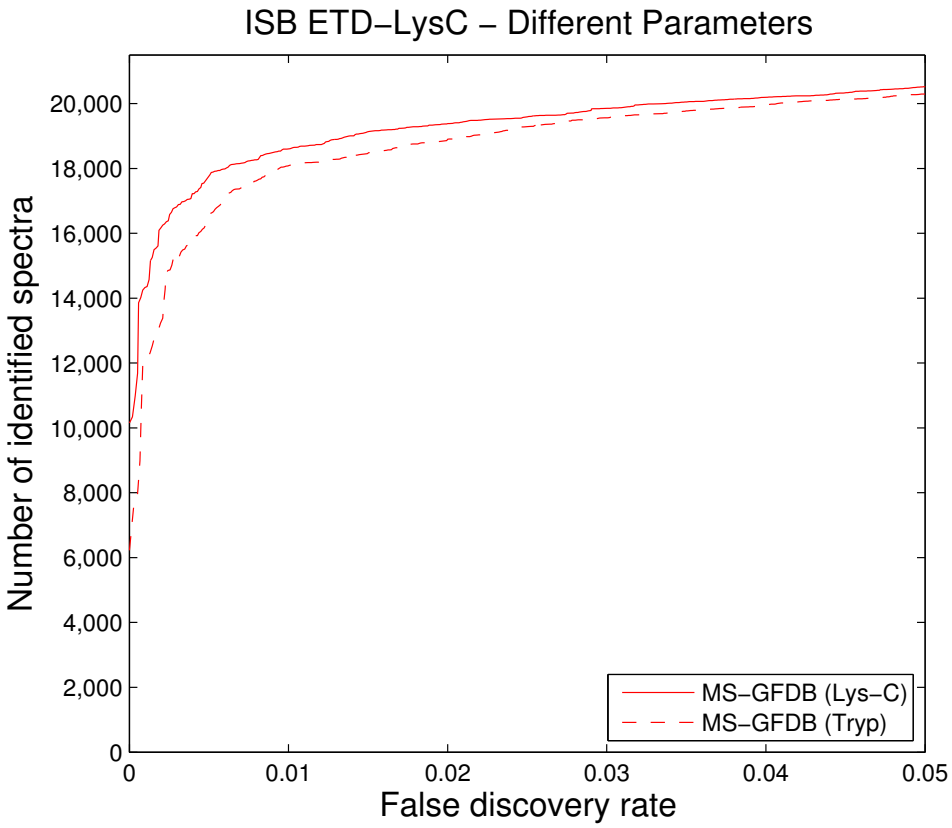
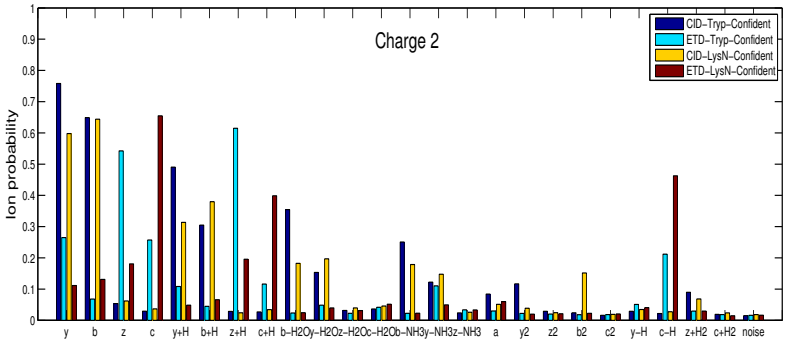
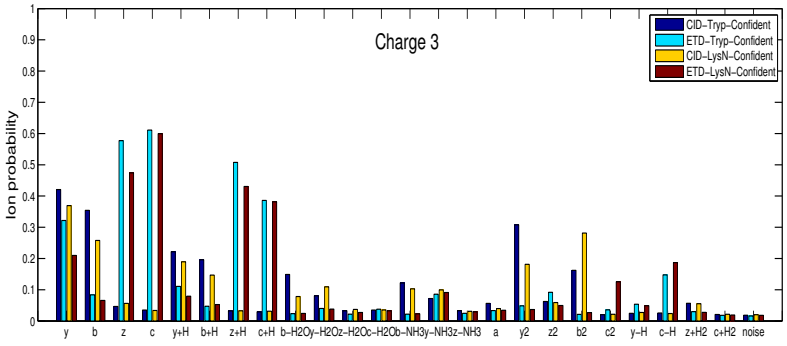


Figure 5.7: Number of identified spectra in the ISB Lys-C data set with MS-GFDB, using the scoring parameters derived from the ISB Lys-C data set from the Coon lab (solid line) and ETD-Tryp data set from the Heck lab (dashed line). The specific choice of the training data set does not significantly affect the MS-GFDB results



(a)



(b)

Figure 5.8: Probabilities of various ion types for the four types of (a) charge 2 spectra and (b) charge 3 spectra (see [33] for similar analysis). Spectra in CID-Tryp-Confident, ETD-Tryp-Confident, CID-LysN-Confident and ETD-LysN-Confident were used. All the spectra were filtered to remove noisy peaks as follows: given a peak at mass M, we retained the peak if it is among the top six peaks within a window of size 100 Da around M. Precursor ions (or charge-reduced precursor ions) and their derivatives were also filtered out. A colored bar represents the probability (y-axis) of a certain type of ion (x-axis) being present in a filtered spectrum. Each data set is color coded. For example, a charge 2 spectrum in CID-Tryp-Confident generated from a length 10 peptide is expected to have 10-1 (number of potential cleavage sites) $\times$ 0.76 (probability of y ion) = 6.8 y ions, while a charge 2 spectrum in ETD-Tryp-Confident is expected to have only 9 $\times$ 0.26 = 2.3 y ions. In MS-GFDB, all ion types with probabilities exceeding 0.15 are used for scoring.

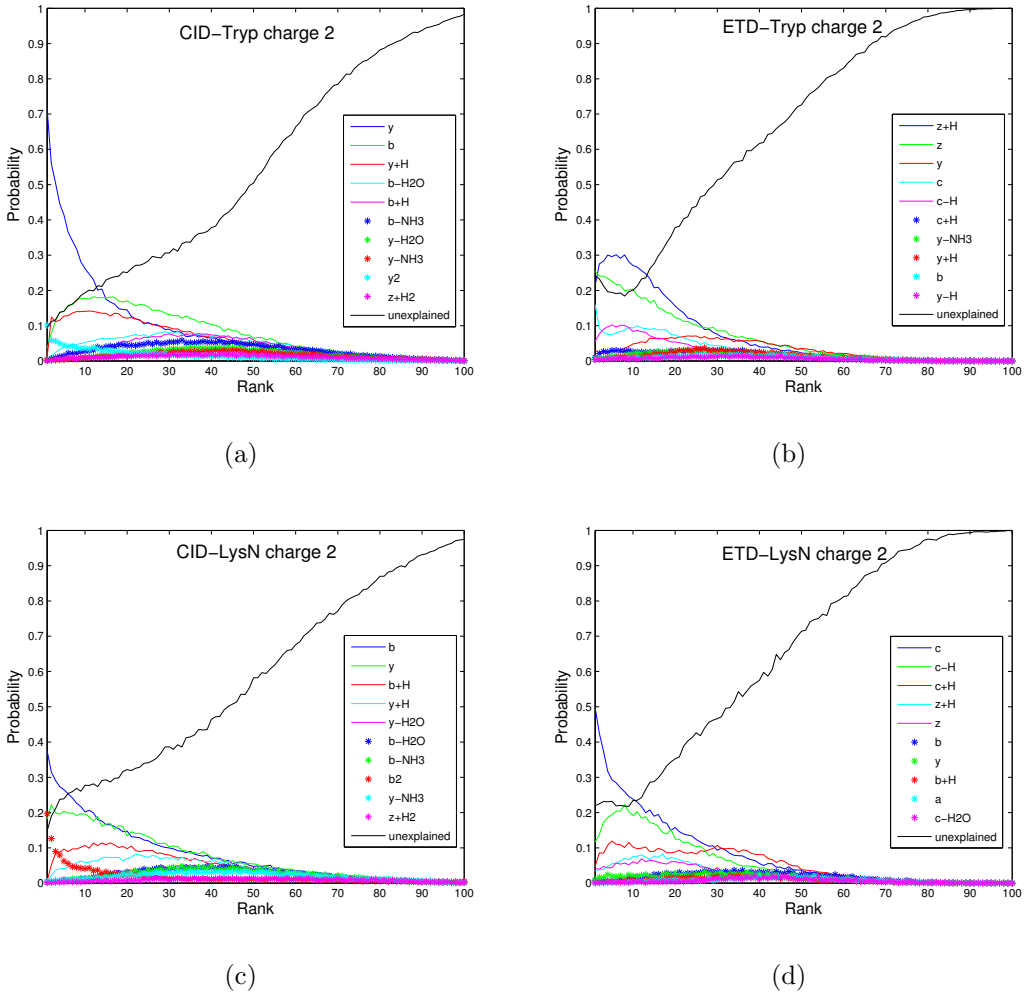


Figure 5.9: Rank distributions of different ion types for different data sets: (a) CID-Tryp-Confident, (b) CID-LysN-Confident, (c) ETD-Tryp-Confident and (d) ETD-LysN-Confident. Only charge 2 spectra were considered and all spectra were filtered to remove precursor ions (or charge-reduced precursor ions) and their derivatives. For each data set, 10 different ion types with highest probabilities were selected and the probability of a peak of a given rank (x-axis) being a certain ion type (color-coded) is plotted for peaks with rank 1 to 100. The black curve (labeled as unexplained) represents the peaks that are not explained by any of the 10 selected ion types. For example, for CID-Tryp-Confident charge 2, the highest ranked peak represents a singly charged y ion with probability 0.7, a doubly charged y ion (y2) with probability 0.1, a singly charged b ion with probability 0.04, etc. It remains unexplained with probability 0.1.

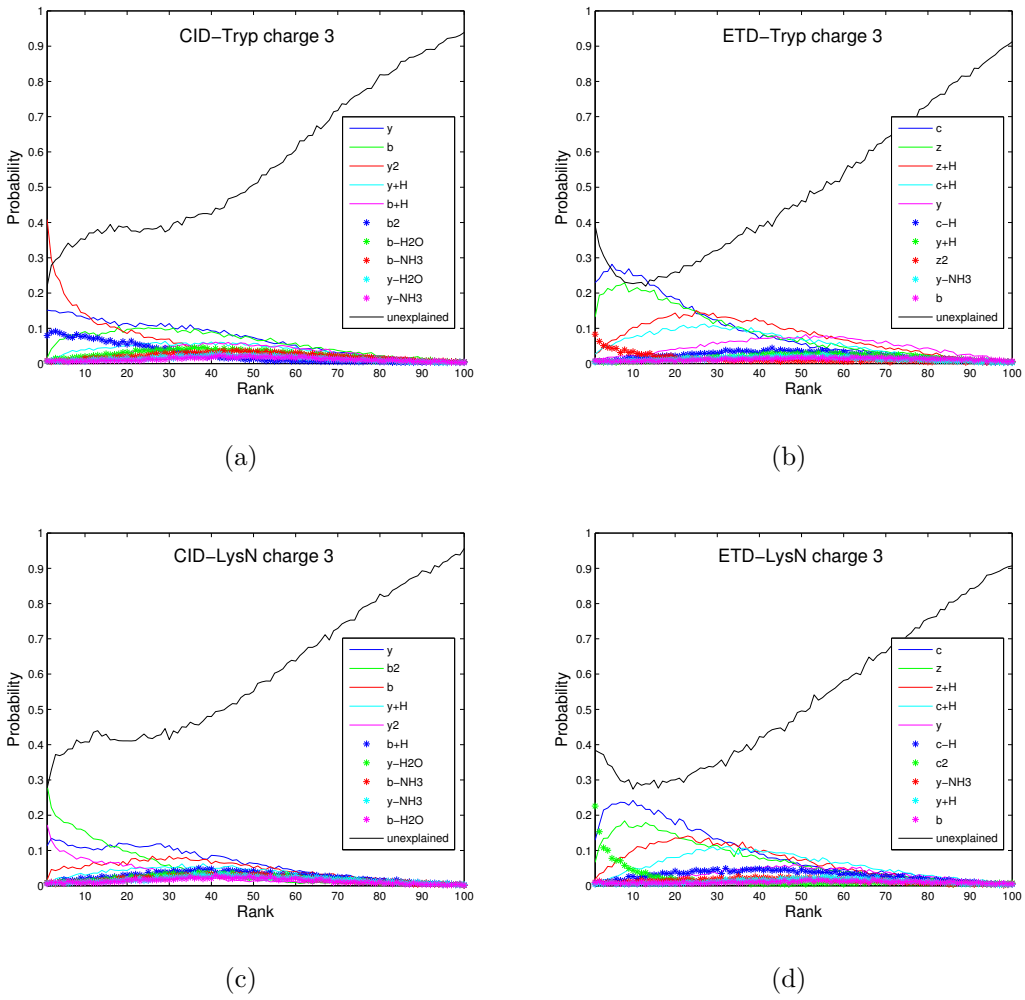
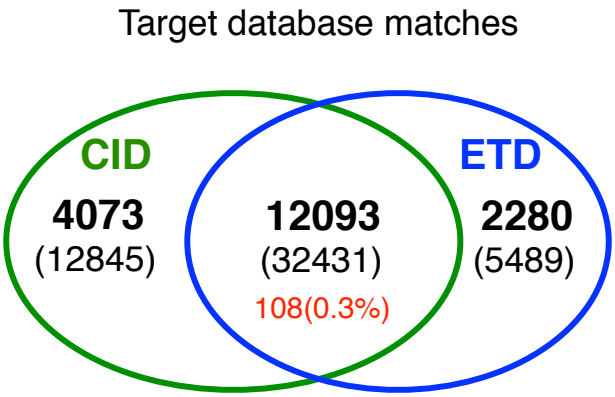
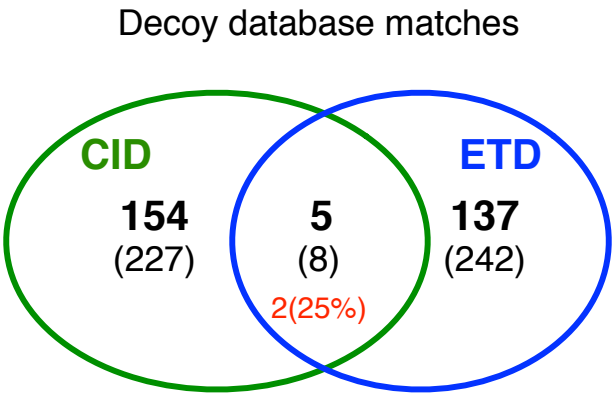


Figure 5.10: Analog of Figure 5.9 for charge 3 spectra. The rank distributions for charge 3 spectra are different from those for charge 2 spectra.



(a)



(b)

Figure 5.11: Venn diagrams of (a) spectral pairs identified against the IPI-Human database within peptide level FDR 1% and (b) spectral pairs identified against the decoy database with p-values corresponding to peptide level FDR 1% or less. The number of peptides (the number of spectral pairs in parentheses) are shown. Numbers in boxes correspond to the number (percentage in parentheses) of spectral pairs where CID and ETD identifications disagree.

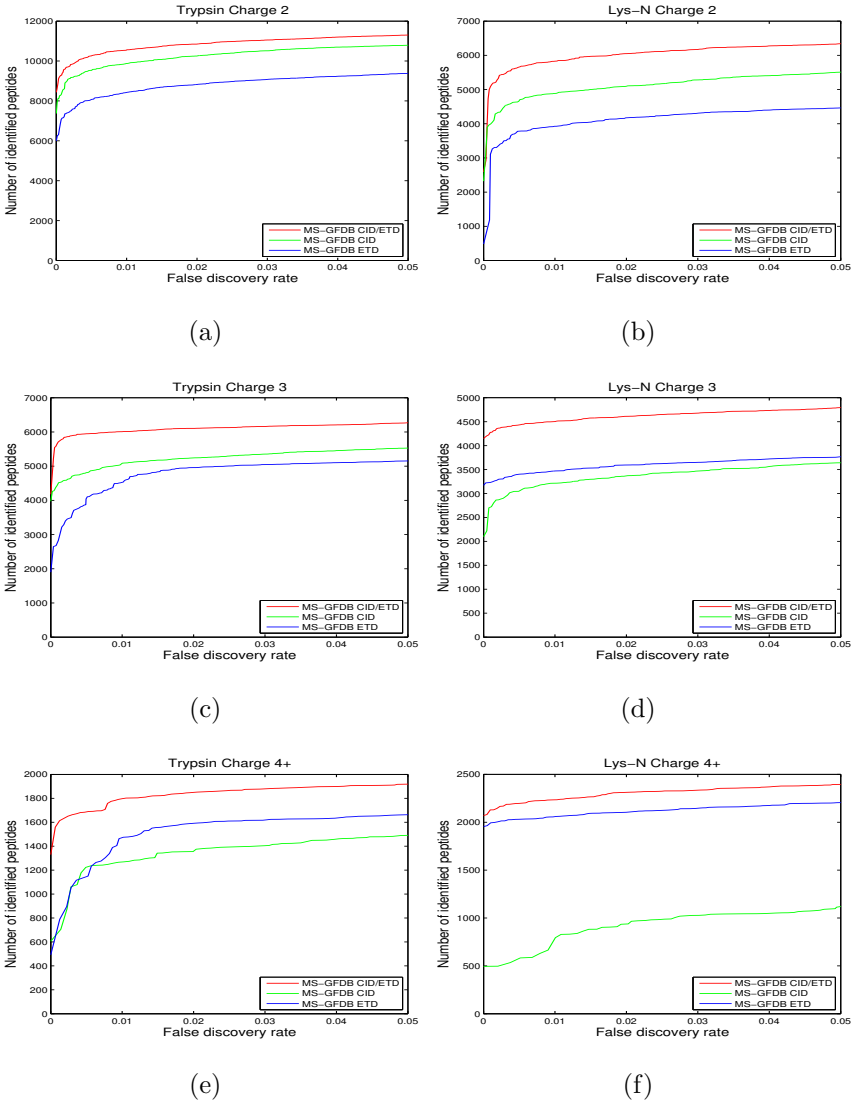


Figure 5.12: Number of identified peptides with MS-GFDB CID/ETD from (a) charge 2 spectral pairs in CID-Tryp and ETD-Tryp, (b) charge 2 spectral pairs in CID-LysN and ETD-LysN, (c) charge 3 spectral pairs in CID-Tryp and ETD-Tryp, (d) charge 3 spectral pairs in CID-LysN and ETD-LysN, (e) spectral pairs of charges 4 and larger in CID-Tryp and ETD-Tryp, and (f) spectral pairs of charges 4 and larger in CID-LysN and ETD-LysN. Number of identified peptides with MS-GFDB are also shown for reference. The number of peptide identifications is plotted against the corresponding peptide level FDR. FDRs were separately computed for spectra of precursor charge 2, precursor charge 3 and precursor charge 4 and larger. Red curves represent MS-GFDB CID/ETD, green curves represent MS-GFDB CID and blue curves represent MS-GFDB ETD. For all the cases considered, MS-GFDB outperformed both MS-GFDB CID and MS-GFDB ETD.

Chapter 6

Universal and Sensitive Database Search Tool

6.1 Introduction

MS instruments and experimental protocols have greatly advanced over the last decade. Several new fragmentation technologies like ETD [104] and Higher-energy Collisional Dissociation (HCD) [142] emerged and high-precision mass spectrometers like Orbitrap and FT-ICR became widely available. While trypsin remains a dominant protease in proteomics studies, digesting proteins with diverse proteases is becoming popular [143]. Empowered by these changes, MS researchers now have diverse choices with respect to the questions: "what fragmentation method to use?", "how accurate should be the measurements of the mass-to-charge (m/z) ratios?", "what proteases to use?", and "what PTM to focus on (e.g. phosphorylation)?". Depending on these choices, the resulting tandem mass (MS/MS) spectra vary in fragmentation propensities and precision. Therefore, unlike in the past when low-precision CID spectra of tryptic peptides dominated the field, spectral datasets generated today are very diverse.

The key to interpreting MS/MS spectra is how to score a PSM. There are two types of approaches to scoring a PSM formed by a peptide P and a spectrum S. The first is to compare S against a "theoretical" spectrum predicted from P using

pre-defined rules (database scoring) [24]. The second is to compare S against a spectrum previously identified as being generated from P if this spectrum is available (library scoring) [7]. This library scoring yields more discriminative scores than the database scoring, since predicting theoretical spectra remains difficult. However, in contrast to the database scoring that can be used to identify *any* peptide, the library scoring is applicable only to identify *some* peptides contained in the spectral library, a collection of previously identified PSMs. In this chapter, we propose a database scoring algorithm having better discriminating power as compared to existing library scoring algorithms.

Since existing spectral libraries are still incomplete, database search is the most commonly used approach to interpret spectra. Unfortunately, existing MS/MS database search tools such as SEQUEST [24] and Mascot [5] have not kept pace with the increased diversity of the data because they are largely optimized for low-precision CID spectra of tryptic peptides [1]. Many efforts have been invested into making MS/MS search tools compatible with new types of data. For example, several pre- or post-processing strategies have been proposed [144, 145], resulting in small improvement in the performance of database search tools. To further boost the performance, MS/MS database search tools are combined with statistical modeling tools like PeptideProphet [11], and Percolator [131]. These tools do not find new Peptide-Spectrum Matches (PSMs), but rather re-score PSMs reported by a database search tool using more complex scoring and output high-scoring PSMs. While they often improve the performance of a database search tool, their performance is negatively affected when the database search tool fails to find correct PSMs [146]. Another downside of the pre- or post-processing strategies and statistical modeling tools is that since they are usually not integrated into database search tools, using them complicates the analysis of MS/MS spectra. Moreover, since different laboratories employ different combinations of tools (see Figure 6.1), even for the same data, the capabilities of analyzing the data vary widely and the results obtained in one laboratory are practically impossible to reproduce in another laboratory.

We advocate using *universal* database search tools that perform well for

diverse types of spectral datasets. To address this need, we developed MS-GF+, a universal database search tool that works well (i.e., identifies more peptides than all other MS/MS tools we tested) for spectra generated using diverse configurations of MS instruments and experimental protocols. We emphasize that MS-GF+ is not customized for specific spectral datasets but rather uses a robust probabilistic model that works well across all datasets.

MS-GF+ is universal because it automatically derives scoring parameters from thousands of PSMs without prior knowledge of the type of the spectra [146]. We represent various types of spectra as a graph where paths represent *spectral types* (Figure 6.1). For each spectral type, MS-GF+ further divides the PSMs into subgroups depending on the precursor ion charge and m/z. Afterwards, it learns scoring parameters separately for each path and each subgroup. To score a PSM, MS-GF+ uses a different set of scoring parameters depending on the spectral type and the subgroup.

MS-GF+ can train the scoring parameters for any spectral type (including spectral types not specified in Figure 6.1) or use the pre-trained scoring parameters. MS-GF+ provides an user interface, taking over the authority to train scoring parameters to the users and making training as easy as running a database search. This is useful to the researchers who use novel MS instruments or experimental protocols. Even researchers using a common configuration (e.g. CID of trypsin digests) will be able to benefit from this function, because spectra generated in different laboratories may vary.

MS-GF+ addresses the following limitations of most existing MS/MS tools (including MS-GF [78] and MS-GFDB [146]): inability to estimate E-values accurately for PSMs formed by modified peptides and limited ability to take advantage of accurate m/z values in high-precision MS/MS spectra. In addition, MS-GF+ greatly reduces running time (as compared to MS-GFDB, see Figure 6.2), and features an improved usability due to ProteoSAFe, a user-frendly interface for searches, reports and data management.[1]

[1] To-ju Huang, Claudiu Farcas, Jeremy Carver, Natalie Castellana, Ari Frank, Sangtae Kim, Jian Wang, Pavel A. Pevzner, Vineet Bafna, Ingolf Krüger, and Nuno Bandeira. ProteoSAFe: A Scalable, Accessible, and Flexible Software Environment for Proteomics Analysis, *in preparation.*

We demonstrate the performance of MS-GF+ using various datasets: spectra of tryptic peptides generated using CID, HCD and ETD in combination with either linear ion trap or Orbitrap readout; spectra of multiple enzyme digests; spectra of phosphopeptides; and spectra or a novel protease alpha-lytic protease (αLP). For all these datasets, we show that MS-GF+ greatly outperforms state-of-the-art tools for peptide identification.

6.2 Methods

Mass spectrometers are usually divided into High-precision (denoted by H) and Low-precision (denoted by L) instruments. Depending on whether the precursor and product ions are measured with Low or High-precision, the spectra are divided into LL, LH, HL, and HH spectra (LH spectra are hardly ever used in proteomics studies).

MS-GF+ takes a spectral dataset *Spectra* and a protein database *Protein DB* as an input and outputs a set of scored PSMs along with statistical significance estimates.[2] The workflow of MS-GF+ comprises the following 4 steps: generating PRM spectra, searching a protein database, computing E-values of PSMs, and estimating FDRs. Below we describe each step as well as how MS-GF takes advantage of HH spectra.

6.2.1 Generating PRM spectra

Rescaling mass values

Transformation of experimental spectra into vectors (PRM spectra) requires *binning*. To support this transformation, MS-GF+ is designed under the assumption that amino acid masses are integers. MS-GF+ uses nominal masses as integer masses of amino acids. While this enables efficient computing of MS-GF+ scores, it causes rounding errors because peaks in the spectrum correspond to *real* rather

[2] *ProteinDB* can be regarded as a long string generated by concatenating all protein sequences (with delimiters).

than *nominal* masses of amino acid sequences. To minimize the rounding errors, MS-GF+ *rescales* every mass m into $0.9995m$ [3, 62, 46]. This dramatically reduces the rounding errors (Table 6.5), so one can estimate the nominal mass of a peptide (or a peak) of mass m by simply taking $[0.9995m]$ where $[x]$ represents the closest integer to x. For example, for a peptide "RESCAILINGMASSES" of mass 1705.776, $[1705.776 \cdot 0.9995 = 1704.92] = 1705$ represents the correct nominal mass of the peptide. We investigated all 188 million unique peptides of length up to 20 in the human IPI database (version 3.87), and the estimation was inaccurate only for 874 peptides. Even if we consider peptides of length up to 40, only 0.07% had estimation errors.

Peptide variant

A (non-modified) *peptide* is defined as a string over the alphabet $\mathcal{A}$ of 20 standard amino acids. MS-GF+ is a restrictive MS/MS database search tool that allows a user to specify a set of allowed modifications. Let $\mathcal{A}^+$ be an *extended* amino acid set containing both unmodified and modified amino acids. For an (unmodified) amino acid $a \in \mathcal{A}$, let $\mathrm{Mod}(a) \subset \mathcal{A}^+$ be the *set* of both unmodified and modified amino acids associated with a. For example, if T (Thr) and T^* (phosphorylated Thr) are in $\mathcal{A}^+$, $\mathrm{Mod}(T) = \{T, T^*\}$. Given a peptide $P = a_1 \dots a_k$, define $PV = pv_1 \dots pv_k$ as a *variant* of P if $pv_i \in \mathrm{Mod}(a_i)$ for all i $(1 \leq i \leq k)$. Throughout this chapter, we abbreviate the peptide variant as *variant*.

Incorporating ion types and ranks

Many database search tools convert a peptide into a "theoretical" spectrum and compare it with an experimental spectrum using scoring functions of various complexity. MS-GF+ converts spectra into *PRM spectra* [146, 25]. A PRM spectrum of a spectrum S is an M-dimensional vector with integer values where M is the nominal parent mass of S.[3] We denote the nominal parent mass

[3] Here, we consider nominal parent masses representing the sum of nominal masses of amino acids of the peptide generated the spectrum. Since in many cases, the precise nominal parent mass is unknown (e.g. MS instruments often choose 2nd or 3rd isotope peak instead of mono-isotope peak from MS1 spectrum), multiple PRM spectra are generated separately for each

of S as ParentMass(S). The conversion from an experimental spectrum to a PRM spectrum proceeds as follows.

A *spectrum* $S = \{(mz_1, rank_1), \ldots, (mz_l, rank_l)\}$ is represented as a set of *ranked* peaks where the ith highest intensity peak gets rank i (mz_j and $rank_j$ represent m/z and rank of jth peak, respectively). An *ion type* is represented as a triplet of integers *charge, offset*, and *sign* representing whether the ion type is a prefix ion ($sign = 1$) or a suffix ion ($sign = -1$). For example, the singly-charged b-ion and y-ion correspond to ion types (1, 1, 1) and (1, 19, -1), respectively. Given an ion type $ion = (charge, offset, sign)$, one can turn a spectrum S of nominal parent mass M into $S_{ion} = \{(prm_1, rs_1), \ldots, (prm_l, rs_l)\}$ using the following transformation:

$$prm_j = \begin{cases} [mz_j \cdot charge \cdot 0.9995] - offset & \text{if } sign = 1 \\ M - \big([mz_j \cdot charge \cdot 0.9995] - offset\big) & \text{if } sign = -1 \end{cases}$$

$$rs_j = \text{RankScore}(ion, rank_j),$$

where RankScore($ion, rank$) is a pre-computed function that takes an ion type ion and an integer $rank$ and returns a probabilistic log-likelihood score defined in [146, 3].[4] Assume that $\mathcal{I}$ is a set of ion types contributing to scoring. The PRM spectrum of S (denoted by PRM(S) $= s_1 \ldots s_M$) is computed as follows:

$$s_i = \sum_{ion \in \mathcal{I}} \max(\{rs \mid (prm, rs) \in S_{ion} \text{ and } prm = i\} \cup \text{RankScore}(ion, \infty)),$$

where RankScore(ion, ∞) represents the score given when ion is missing.

We also define a PRM spectrum of a variant as follows. Let Mass(a) be the nominal mass of a (possibly modified) amino acid a. For example, Mass(T) = 101 and the mass of phosphorylated Thr is Mass(T^*) = 181. Given a variant $PV = pv_1 \ldots pv_k$, define the mass of PV as $\text{Mass}(PV) = \sum_{i=1}^{k} \text{Mass}(pv_i)$. Given a variant $PV = pv_1 \ldots pv_k$ of mass M, we define its PRM spectrum (denoted by PRM(PV)) as a 0-1 vector $m_1 \ldots m_M$ with $(n-1)$ 1s, such that $m_i = 1$ if i equals to Mass(pv_1) + ... + $Mass(pv_j)$ ($1 \le j \le k$).

possible nominal parent mass, and the score of a peptide of mass M is computed from the PRM spectrum of parent mass M.

[4] In practice, RankScore($ion, rank$) also accounts for the location of the observed peak and the precursor charge and mass of the spectrum, which are omitted here for simplification.

Removing precursor-related peaks

Some types of spectra (e.g., ETD spectra) often feature high-intensity precursor peaks, charge-reduced precursor peaks, and their neutral and side-chain losses. While these peaks are not informative, they may artificially inflate scores of false PSMs if they fall in the positions of certain ion types. MS-GF+ removes precursor-related peaks prior to generating PRM spectra to eliminate a risk of erroneously interpreting them as other ion types. The information on which peaks to remove is automatically pre-computed from a training set [146].

Scoring a peptide-spectrum match

The *MS-GF+ score* of a PSM (PV, S) is defined as $\text{MSGFScore}(PV, S) = \text{PRM}(PV) \cdot \text{PRM}(S)$ if $\text{Mass}(PV) = \text{ParentMass}(S)$ and $-\infty$ otherwise. The MS-GF+ score represents the log likelihood ratio described in [3].

6.2.2 Searching a protein database

We define $ProteinDB^+$ as the set of all variants (with respect to an extended amino acid set $\mathcal{A}^+$) derived from $ProteinDB$. The goal of MS-GF+ database search is to solve the following database search problem:

Database search problem. Given a spectral dataset $Spectra$ and a protein database $ProteinDB$, for each spectrum $S \in Spectra$ find a variant $PV_{S,ProteinDB}$ such that

$$PV_{S,ProteinDB} = \underset{PV \in ProteinDB^+}{\arg\max} \text{MSGFScore}(PV, S).$$

Solving this problem involves the following three steps: (1) for every spectrum $S \in Spectra$, computing $\text{PRM}(S)$, (2) for every variant $PV \in ProteinDB^+$, computing $\text{PRM}(PV)$, and (3) for every pair of (PV, S) where $\text{Mass}(PV) = \text{ParentMass}(S)$, computing $\text{MSGFScore}(PV, S) = \text{PRM}(PV) \cdot \text{PRM}(S)$. To execute these steps efficiently, one may simply execute the step (1) and (2), store all $\text{PRM}(S)$ and $\text{PRM}(PV)$ in the main memory and execute the step (3). But this is often infeasible because the number of variants is usually too large to fit all $\text{PRM}(PV)$ in the main

memory. Alternatively, one may consider executing the step (2) on the spot for each spectrum, but this is prohibitively slow.[5]

Instead of storing both PRM(S) and PRM(PV), MS-GF+ stores only PRM(S) for all spectra in the main memory, and indexes them by parent masses. Since PRM spectra compactly represent experimental spectra, MS-GF+ can store over 200,000 PRM spectra in the main memory of 4GB. Rather than finding the best scoring peptide for each spectrum, MS-GF+ then finds the best scoring spectrum for each variant. Formally, MS-GF+ solves a slightly different problem: for each variant $PV \in ProteinDB^{+}$, find a spectrum $S_{PV,Spectra}$ such that

$$S_{PV,Spectra} = \underset{S \in Spectra}{\arg\max}\, \text{MSGFScore}(PV, S) = \underset{S \in Spectra_{\text{Mass}(PV)}}{\arg\max}\, \text{MSGFScore}(PV, S), \tag{6.1}$$

where $Spectra_{\text{Mass}(PV)}$ represents the set of spectra S with ParentMass(S) equals to Mass(PV). This problem can be solved efficiently by enumerating variants $PV \in ProteinDB^{+}$ one by one, generating PRM(PV) *on the spot*, and computing MSGFScore(PV, S) = PRM(PV) $\cdot$ PRM(S) for all *pre-computed* PRM(S) where $S \in S_{\text{Mass}(PV)}$. Once $S_{PV,Spectra}$ is found for all variants PV, finding $PV_{S,ProteinDB}$ is trivial because

$$PV_{S,ProteinDB} = \underset{PV \in \{PV | S_{PV,Spectra} = S, PV \in ProteinDB\}}{\arg\max}\, \text{MSGFScore}(PV, S).$$

In practice, to save the memory, instead of recording $S_{PV,Spectra}$ for every PV, MS-GF+ records the best scoring variant PV^{*} while enumerating PVs, and updates PV^{*} to PV whenever it finds PV where $S_{PV,Spectra} = S$ and MSGFScore(PV, S) > MSGFScore(PV^{*}, S).

Similar to pFind [147], MS-GF+ uses a suffix array (a lexicographically sorted list of all the suffixes of $ProteinDB$ [148]) to further optimize the database search. Protein databases (particularly, eukaryotic ones) contain many similar proteins, so many peptides appear in multiple copies in a database (*repeated peptides*). For example, the IPI human database (version 3.87) contains about 130,000 fully-tryptic peptides of length 10, but the number decreases to about 50,000 if only

[5]For the IPI human database, executing the step (2) alone takes 54 seconds, considering partially tryptic peptides of lengths between 6 and 40, and two variable modifications Oxidation of Met and protein N-term Acetylation.

unique peptides are considered. If a protein database is indexed as a suffix array, peptide occurrences from the same repeated peptide appear in neighboring indices in the suffix array. So, instead of searching peptides according to their ordering in the original database file, MS-GF+ searches peptides according to their ordering in the suffix array, and uses the longest common prefix data structure [148] to score each unique peptide only once (Figure 6.2 (b)).

Pre-filtering of low-quality spectra

Since some MS/MS spectra feature limited fragmentation (or excessive number of unexplained peaks), they are unlikely to be identified by MS-GF+ (or most other database search tools). MS-GF+ filters out such *low-quality* spectra *prior to* searching a database, leading to a reduction in running time. As shown by Frank et al. [90], such filtration also increases the number of identified PSMs at a fixed FDR (since the number of spurious PSMs decreases). Existing methods [90, 149, 150] often use machine learning techniques to separate high and low-quality spectra. In addition to computing $\arg\max_{PV \in ProteinDB^+} \text{MSGFScore}(PV, S)$, MS-GF+ also has capability to compute $\text{DeNovoScore}(S) = \max_{PV} \text{Score}(PV, S)$, where maximum is taken over all variants rather than over $PV \in ProteinDB^+$ as before. Since DeNovoScore represents the log-likelihood ratio, spectra S with negative DeNovoScore(S) are likely to represent spurious PSMs, and can be discarded. Note that DeNovoScore(S) can be computed using fast linear-time dynamic programming algorithms [33].

Comparison with MS-GFDB

MS-GF+ achieved an order of magnitude speedup compared to MS-GFDB. Furthermore, pre-processing of the database required to run MS-GF+ is much faster as compared to MS-GFDB which uses indices from peptide masses to the peptide locations (Figure 6.2). The indexing in MS-GFDB had to be repeated even for the same protein database depending on the chosen enzymes and allowed amino acid modifications. In contrast, MS-GF+ uses a suffix array of a protein database that needs to be constructed only once, i.e., it does not depend on the

chosen enzyme and allowed modifications. Also suffix arrays in MS-GF+ can be constructed much faster than indices in MS-GFDB while being more memory-efficient [147].

6.2.3 Computing E-values

Given a spectrum S, a score threshold t, an extended set of amino acids $\mathcal{A}^+$, and a database size N, we define E-value$(S, t, \mathcal{A}^+, N)$ as the expected number of variants PV (as defined by $\mathcal{A}^+$) with MSGFScore$(PV, S) \geq t$ in a random protein database of size N. To compute E-value$(S, t, \mathcal{A}^+, N)$, we first compute *spectral E-value* E-value$(S, t, \mathcal{A}^+)$, the expected number of variants PV with MSGFScore$(PV, S) \geq t$ given a *single random peptide*. A single random peptide models a random peptide starting at a fixed position in a random protein database.

We consider a set of all possible (unmodified) peptides of length k (where k is a large number) and select a random peptide uniformly from this set (i.e. the probability of selecting a peptide is $\frac{1}{20^k}$).[6] We say that a peptide P *produces* a variant PV if PV is a variant of a prefix of P. For example, $PEPT^*$ and $PEPTI$ are produced by $PEPTIDE$. Given a spectrum S, let $\mathcal{PV}(t)$ be the set of all variants PV with MSGFScore$(PV, S) \geq t$. For every variant PV, there are $20^{k-|PV|}$ peptides of length k producing a variant PV ($|PV|$ stands for the number of amino acids in PV). Therefore, expected number of variants per random peptide with a score equal or better than t is

$$\text{E-value}(S, \mathcal{A}^+, t) = \sum_{PV \in \mathcal{PV}(t)} \frac{20^{k-|PV|}}{20^k} = \sum_{PV \in \mathcal{PV}(t)} 20^{-|PV|}.$$

Since a variant is a string over the alphabet $\mathcal{A}^+$, this expression can be computed using the generating function approach [78]. Given a spectrum S with PRM$(S) = s_1 \dots s_M$, consider a direct acyclic graph called an *amino acid graph* $G(V, E, \mathcal{A}^+)$ with $V = \{0, \dots, M\}$ and $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$, where the *score* of a vertex i is defined as s_i, the *probability* of an edge is defined as $\frac{1}{20}$, the

[6] In practice, to reflect different frequencies of amino acids in a database (e.g. Leu is usually more common than Trp), we define the probability of a peptide $P = a_1 \dots a_k$ as $\prod_{i=1}^{k} \text{Prob}(a_i)$ where Prob(a) is the frequency of amino acid a in a protein database. Note that this does not change the algorithm to compute spectral E-values.

score of a path is defined as the sum of scores of its vertices, and the *probability* of a path is defined as the product of probabilities of its edges. A path in an amino acid graph represents a variant. Therefore, E-value$(S, \mathcal{A}^+, t)$ equals to the sum of probabilities of all paths from 0 to M with scores equal or better than t, and can be computed using parametric dynamic programming [3, 78, 151].

While spectral E-values are useful for evaluating statistical significance of individual PSMs (independently of the database), they need to be transformed into E-value$(S, t, \mathcal{A}^+, N)$ to take into account the fact that the database search represents "multiple testing" where multiple variants (arising from different database peptides) are scored against a spectrum [152]. E-values can be approximated as follows:

$$\text{E-value}(S, t, \mathcal{A}^+, N) = \text{E-Value}(S, t, \mathcal{A}^+) \cdot N,$$

where N is the size of the database.[7]

6.2.4 How to benefit from high-precision MS/MS spectra?

While it may appear that addressing all LL, HL, and HH spectra is a simple matter of tuning parameters that control the error tolerance, the situation is more complex. Here we explain how MS-GF+ takes advantage of high-precision product ion peaks.

Database search of HL and HH spectra

Let RMass(a) be the real mass of an amino acid a. For a variant $PV = pv_1 \dots pv_k$, let RMass$(PV) = \sum_{i=1}^{k}$ RMass(pv_i), and RParentMass(S) be the real parent mass of a spectrum S. We previously defined MSGFScore$(PV, S) =$ PRM (PV) ·PRM(S) if Mass$(PV) =$ ParentMass(S) and $-\infty$ otherwise. Note that the condition Mass$(PV) =$ ParentMass(S) is weak, i.e., it may be satisfied even when real mass RMass(PV) significantly deviates from RParentMass(S). To take advantage of accurate parent masses in HL and HH spectra, this condition has to

[7]Since protein databases contain many repeated peptides in practice, it is important to reflect an *effective* size of the database that is estimated as the number of unique peptides of certain length.

be redefined to $\text{RMass}(PV) - \Delta < \text{RParentMass}(S) < \text{RMass}(PV) + \Delta$, where Δ is the precursor mass tolerance. To solve the database search problem for this modified definition of MSGFScore, the equation 6.1 should be changed as follows:

$$S_{PV,Spectra} = \arg\max_{S \in Spectra} \text{MSGFScore}(PV, S) = \arg\max_{S \in Spectra_{\text{RMass}(PV)}} \text{MSGFScore}(PV, S), \tag{6.2}$$

where $Spectra_{\text{RMass}(PV)}$ represents the set of spectra $S \in Spectra$ satisfying

$$\text{RMass}(PV) - \Delta < \text{RParentMass}(S) < \text{RMass}(PV) + \Delta.$$

Scoring HH spectra

In [3], we introduced an abstract model (seemingly unrelated to mass spectrometry) and described a probabilistic process of transforming a Boolean string into another Boolean string. Below, we describe a transformation of a Boolean string into a directed acyclic graph (DAG) and generalizes this model for scoring real PSMs.

Let $P = p_1 \dots p_M$ be a Boolean string called a *peptide*. Let $G_S = (V, E, \mathcal{A}^+)$ be a labeled DAG called a *G-spectrum* with vertices $V = \{0, \dots, M\}$, and edges $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$. We define the Boolean label of vertex i as v_i and edge (i, j) as $e_{i,j}$. The probability of P generating G_S is defined as follows:

$$\text{Prob}(G_S|P) = \prod_{i \in V} \text{Prob}_{\text{V}}(v_i|p_i) \cdot \prod_{(i,j) \in E} \text{Prob}_{\text{E}}(e_{i,j}|p_i, p_j),$$

where $\text{Prob}_{\text{V}}(x|y)$ is a 2×2 matrix representing the probability of a peptide character (0 or 1) generating a vertex label, and $\text{Prob}_{\text{E}}(x|y, z)$ is a 2×4 matrix representing the probability of a pair of peptide characters generating an edge label (Figure 6.3). In practice, $\beta_1 \approx \beta_2 \approx \beta_3$.

When applying this model for scoring a peptide P and a G-spectrum G_S, we consider a test comparing two hypotheses: one assuming G_S is generated by P and the other assuming G_S is generated by a string consisting of all zeros (denoted by O). The score of (P, G_S) (denoted by $\text{Score}(P, G_S)$) is defined as follows (see

Figure 6.4 for an example):

$$
\begin{aligned}
\mathrm{Score}(P, G_S) &= \log \frac{\mathrm{Prob}(G_S|P)}{\mathrm{Prob}(G_S|O)} \\
&= \log \frac{\prod_{i\in V} \mathrm{Prob_V}(v_i|p_i) \cdot \prod_{(i,j)\in E} \mathrm{Prob_E}(e_{i,j}|p_i, p_j)}{\prod_{i\in V} \mathrm{Prob_V}(v_i|0) \cdot \prod_{(i,j)\in E} \mathrm{Prob_E}(e_{i,j}|0,0)} \\
&= \sum_{i\in V} \log \frac{\mathrm{Prob_V}(v_i|p_i)}{\mathrm{Prob_V}(v_i|0)} + \sum_{(i,j)\in E} \log \frac{\mathrm{Prob_E}(e_{i,j}|p_i, p_j)}{\mathrm{Prob_E}(e_{i,j}|0,0)} \\
&\approx \underbrace{\sum_{i\in\{i|i\in V, p_i=1\}} \underbrace{\log \frac{\mathrm{Prob_V}(v_i|1)}{\mathrm{Prob_V}(v_i|0)}}_{\mathrm{VertexScore}(i)}}_{\text{vertex scoring}} \\
&\quad + \underbrace{\sum_{(i,j)\in\{(i,j)|(i,j)\in E, p_i=1, p_j=1\}} \underbrace{\log \frac{\mathrm{Prob_E}(e_{i,j}|1,1)}{\mathrm{Prob_E}(e_{i,j}|0,0)}}_{\mathrm{EdgeScore}(i,j)}}_{\text{edge scoring}}
\end{aligned}
\tag{6.3}
$$

Note that in the last equation, only the edges (i, j) where $p_i = 1$ and $p_j = 1$ contribute to the edge scoring because $\beta_1 \approx \beta_2 \approx \beta_3$.

We now explain how to convert a spectrum S into a G-spectrum given $\mathcal{A}^+$ and $\mathcal{I}$. Vertex and edge sets are constructed as described earlier. For simplicity, suppose that $\mathcal{I} = \{(1, 0, 1)\}$ (i.e. only singly charged prefix ion with an offset zero contributes to the scoring). Given a constant δ called a *fragment mass tolerance*, two peaks of S with m/z x and y form a *duo* if $y - x$ is approximately equal to a mass of an amino acid, i.e., $\mathrm{RMass}(a) - \delta < y - x < \mathrm{RMass}(a) + \delta$ for $a \in \mathcal{A}^+$. The vertex label v_i and the edge label $e_{i,j}$ of G_S are defined as follows: $v_i = 1$ if there exists a peak of mass x satisfying $[0.9995 \cdot x] = i$ and $v_i = 0$ otherwise; $e_{i,j} = 1$, if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x] = i$ and $[0.9995 \cdot y] = j$, and $e_{i,j} = 0$ otherwise.

In practice, we generate multiple *G-spectra* for a single spectrum, one for each $ion \in \mathcal{I}$. To generate a G-spectrum for $ion = (charge, offset, sign)$ with a real offset $roffset$, (e.g. real offset of the singly-charged b-ion is 1.008), we convert $S = \{(mz_1, rank_1), \ldots, (mz_l, rank_l)\}$ into $S' = \{(mass_1, rank_1), \ldots, (mass_l, rank_l)\}$

using the following transformation:

$$mass_j = \begin{cases} mz_j \cdot charge - roffset & \text{if } sign = 1 \\ \text{RParentMass}(S) - \big(mz_j \cdot charge - roffset\big) & \text{if } sign = -1 \end{cases}$$

Each peak of S representing *ion* corresponds to a peak of this *converted spectrum* S' representing an ion type $(1, 0, 1)$. Therefore, the vertex and edge labels of the G-spectrum for *ion* are defined exactly the same as outlined before, but using S' instead of S (Figure 6.5).

In reality, integer instead of Boolean values are used for vertex and edge labels of G-spectra. Given a converted spectrum S', first all peaks $(x, rank)$ are removed if there exists another peak $(x', rank')$ where $[0.9995 \cdot x] = [0.9995 \cdot x']$ and $rank > rank'$. The vertex label v_i is defined as follows: $v_i = rank$ if there exists a peak $(x, rank)$ satisfying $[0.9995 \cdot x] = i$ and $v_i = 0$ otherwise. For an integer m, let AminoAcid(m) be the set of amino acids $a \in \mathcal{A}^+$ satisfying Mass$(a) = m$ (e.g. AminoAcid$(128) = \{\text{Gln}, \text{Lys}\}$). The edge label $e_{i,j}$ is defined as follows: $e_{i,j} = [100 \cdot \min_{a \in \text{AminoAcid}(j-i)}(y - x - \text{RMass}(a))]$ if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x] = i$ and $[0.9995 \cdot y] = j$, and $e_{i,j} = \infty$ otherwise. The constant 100 is multiplied to discretize the real-valued errors into bins of size 0.01 Da.

In this G-spectrum representation, vertex labels encode the information on the *intensities* of individual peaks, and the edge labels encode the information on the *mass errors* of pairs of peaks assuming they represent consecutive peaks of the same ion type. Note that edge labels are independent on peak intensities.

Given a set of G-spectra of S, we generate a *PRM graph* (instead of a PRM spectrum) of S. A PRM graph is a direct acyclic graph with a vertex set $V = \{0, \ldots, M\}$ and and edge set $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$, where the score of a vertex i is the *sum* of VertexScore(i) over all G-spectra of S, the score of an edge (i, j) is the *sum* of EdgeScore(i, j) over all G-spectra of S, the probability of an edge is defined as $\frac{1}{20}$, the score of a path is defined as the sum of scores of its *vertices and edges*, and the probability of a path is defined as the product of probabilities of its edges. Note that a path of a PRM graph represents a peptide (or a variant), and the score of a path represents the score of the peptide

represented by the path.

Given a PRM graph, one can compute spectral E-values of peptides (or variants) using the generating function approach [78]. The generating function approach works for any DAG as long as the score of a path (variant) in the DAG is represented as the sum of scores of the vertices along the path. While the PRM graph has scores both on vertices and edges, it can be converted into a DAG having scores only on vertices by substituting every edge (i, j) into a new vertex ij and two edges (i, ij) and (ij, j). Thus, the generating function approach can be applied to this "PRM graph" scoring model.

In theory, one can apply this PRM graph scoring model to all HH, HL, and LL spectra. However, for HL and LL spectra, using this PRM graph scoring model does not significantly improve over the simpler PRM scoring model because while the running time roughly doubles, the number of peptide identifications only slightly increases (less than 5%). Thus, we apply the PRM graph scoring model only to HH spectra. For HH spectra, we found that it is beneficial to convert multiply charged product ion peaks into singly charged ion peaks (charge deconvolution) prior to generating PRM graphs. We use the following simple algorithm for the charge deconvolution: if two peaks are separated by (mass of $^{13}\mathrm{C}$ − mass of $^{12}\mathrm{C})/c$ within a small tolerance (e.g. 0.01 Da), we assume they are charge c and convert them into charge 1.

Are spectral E-values accurate?

E-values reported by MS-GF+ are accurate for LL spectra, but slightly biased for HL or HH spectra. This is because there is a discrepancy between the search space of the database search and the E-value computation presented in the previous section; in the high-precision setting, peptides considered are those with masses matching the parent mass of the input spectrum within a narrow tolerance (e.g. 10 ppm) (see Equation 6.2); but in the E-value computation, peptides considered are those with masses matching the nominal parent mass of the input spectrum. For HL and HH spectra, MS-GF+ E-values are larger than they should be (conservative estimation) when a strict precursor mass tolerance (e.g. 10ppm)

is used.

6.2.5 Estimating FDRs

Existing MS/MS database search tools output a set of PSMs and estimate the FDR of this set (fraction of erroneous PSMs) using the *Target-Decoy Approach* (TDA). In addition to computing the FDR via TDA, MS-GF+ also provide a possibility to estimate the FDR via E-values of PSMs (denoted by Expected FDR or EFDR) without using TDA [153]. Since MS-GF+ E-value estimates are accurate for LL spectra but conservative for HH or HL spectra, MS-GF+ EFDRs are also accurate for LL spectra but conservative (higher than they should be) for HH and HL spectra (Figure 6.6).

Peptide identifications are usually reported for a fixed FDR either at the PSM-level (FDR for identified PSMs) or the peptide-level (FDR for identified peptides). Since a single peptide often generate multiple spectra, the same PSM- and peptide-level FDR may result in vastly different set of identified PSMs. MS-GF+ reports peptide-level FDRs along with PSM-level FDRs. To compute peptide level FDRs, for PSMs matched to the same peptide, MS-GF+ retains only the PSM with the lowest spectral E-value. The peptide-level FDR is calculated as $N_{decoyPep}/N_{targetPep}$ where $N_{targetPep}$ ($N_{decoyPep}$) is the number of *retained* target (decoy) peptides with spectral E-values equal or smaller than t.

6.3 Results

6.3.1 MS-GF+ Scoring

Database search tools use a scoring function $\text{Score}(P, S)$ to evaluate a PSM of P and S and further compute statistical significance (e.g. E-values) of the resulting PSMs. Let P_S be a peptide that generated S. A scoring function is *adequate* for S (with respect to a protein database $ProteinDB$) if the correct peptide attains the maximal score in the database, i.e., $\max_{P \in ProteinDB} \text{Score}(P, S) = \text{Score}(P_S, S)$. A "good" scoring function should satisfy the following conditions:

(a) it should be adequate for the great majority of spectra,

(b) the algorithm for PSM scoring should be fast,

(c) the algorithm for computing statistical significance of PSMs should be fast and accurate.

MS-GF+ uses a very simple dot-product scoring $Score(P, S) = P^* \cdot S^*$ after converting peptide P and spectrum S into vectors P^* and S^* referred to as Prefix-Residue-Mass (PRM) spectra [146, 25]. Conversion of a spectrum S into a PRM spectrum S^* uses a probabilistic model that ensures that the resulting dot-product scoring is adequate [3] (condition (a)). At the same time, it makes scoring and computing accurate E-values fast [78] (condition (b) and (c)). This "PRM scoring" model contrasts with many other database search tools using sophisticated scoring functions [24, 13, 49, 154] that often make it difficult to satisfy the condition (c).

The scores of PSMs reported by existing MS/MS database search tools are often poorly correlated with their statistical significance (e.g., E-values). It is important to rank PSMs based on their statistical significance, because such ranking (rather than ranking based on "raw scores") often dramatically increases the number of identified spectra [78, 155]. This observation explains why the condition (c) is important. Many database search tools estimate an E-value of a PSM based on an approximation of a tail of the score distribution specific to the spectrum using *peptides in the database* [13, 49, 155]. Since this approach often results in biased estimates of statistical significance [78], MS-GF+ adopted the generating function approach to rigorously compute E-values of PSMs using the score distribution of *all peptides* [78]. The PRM scoring model is essential here, because the generating function approach is applicable only to the scoring functions that can be represented as a dot-product of vectors [153]. Adopting the generating function approach improves both the accuracy of E-value estimates and increases the number of identified peptides.

The key part of the generating function approach is the assumption that amino acids have integer masses (otherwise the parametric dynamic programming is difficult to implement). However, rounding amino acids to integers introduce

rounding errors. These rounding error reduces after rescaling by 0.9995, making it appropriate for LL and HL spectra (Online Method). However, for HH spectra, the rounding error remains too large even after rescaling, prohibiting MS-GF+ from benefiting from precise product ion peaks. A possible solution to this problem is to change the constant used for rescaling. For example, for a scaling constant 274.335215 (e.g. $\text{mass}(G) = 57.021464 * 274.335215 = 15642.995586 \approx 15643$), the rounding error is bounded by 2.5 parts per million (ppm), which is appropriate for analyzing HH spectra. However, since the time complexity of the generating function algorithm is proportional to 1 over the rescaling constant, this makes computing E-values prohibitively slow. Here we present a new scoring algorithm taking advantage of accurate product ion masses while not substantially increasing the running time of MS-GF+ (Online Methods).

6.3.2 Datasets

For benchmarking MS-GF+, we collected publicly available datasets that were previously studied [143, 156, 157] as well as datasets obtained with a novel protease. Overall, we used 18 datasets ($\approx$ 2.38 million spectra from human, yeast, mouse, and *Schizosaccharomyces pombe* reflecting the diversity of MS data, corresponding to 16 distinct spectral types.

Human datasets with varying fragmentations and instruments

Five human datasets corresponding to the spectral types (**CID**,**Low**,Sta ndard,Trypsin), (**CID**,**High**,Standard,Trypsin), (**ETD**,**Low**,Standard,Trypsin), (**ETD**,**High**,Standard,Trypsin), and (**HCD**,**High**,Standard,Trypsin) contain 38, 401, 33,586, 30,451, 25,734, and 37,810 spectra respectively. These datasets are generated in the Heck laboratory (Utrecht University). HEK293 whole cell lysates were digested by trypsin and analyzed by LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen), using combinations of one of the 3 fragmentation modes CID, ETD, and HCD, and either ion trap or Orbitrap readout for product ion m/z. The detailed experimental procedures are described in [156].

Yeast datasets with varying enzymes

Ten yeast datasets corresponding to the spectral types (**CID**,Low,Standard, **Trypsin**), (**CID**,Low,Standard,**LysC**), (**CID**,Low,Standard,**ArgC**), (**CID**,Low, Standard,**GluC**), (**CID**,Low,Standard,**AspN**), (**ETD**,Low,Standard,**Trypsin**), (**ETD**,Low,Standard,**LysC**), (**ETD**,Low,Standard,**ArgC**), (**ETD**,Low,Standard, **GluC**), and (**ETD**,Low,Standard,**AspN**), contain 333,203, 278,336, 114,351, 81, 669, 251,974, 72,463, 246,428, 204,860, 88,403, and 262,635 spectra, respectively. These datasets were generated in the Coon laboratory (University of Wisconsin Madison). Yeast whole cell lysates were digested separately, with either trypsin, LysC, ArgC, GluC, or AspN, separated into 12 fractions via strong cation exchange (SCX) chromatography and analyzed in triplicate with an ETD-enabled LTQ-Orbitrap mass spectrometer, where peptide fragmentation was accomplished either with CID or ETD using the decision-tree acquisition mode [115].[8] We downloaded 180 (5 enzymes $\times$ 12 fractions $\times$ 3 replicates) spectrum files (Thermo RAW format) and converted each raw file into two mgf files one containing CID and the other containing ETD spectra using "msconvert" in ProteoWizard [158] with "no filtering" option. The conversion was unsuccessful for 6 out of 180 files (5 from Arg-C and 1 from Glu-C digests). These 6 files were removed in the further analyses. The detailed experimental procedures are described in [143].

Mouse dataset of phosphopeptides

A mouse dataset corresponding to the spectral type (**CID**,Low,Phosphorylation,**Trypsin**) contains 181,093 spectra. This dataset was generated from the Gygi laboratory (Harvard Medical School). Nine mouse organ proteins were digested with trypsin and the resulting peptides were fractionated via SCX. Phosphopeptides were enriched via immobilized metal affinity chromatography and analyzed in duplicates via LC-MS/MS on an LTQ-Orbitrap mass spectrometer. Out of 9 organ tissues analyzed, we used the spectra generated from the brain tissue. The detailed experimental procedures are described in [157].

[8] In the decision-tree acquisition mode, the mass spectrometer automatically determines the fragmentation method based on the charge of m/z of precursor ions.

Spectra of αLP digests

Two datasets corresponding to the spectral type (**CID**,Low,Standard,α**LP**) and (**ETD**,Low,Standard,α**LP**) contain 49,167 spectra each. These datasets were generated in the Komives laboratory (University of California, San Diego). The detailed experimental procedures to generate these datasets are as follows. Wild-type *S. pombe* cells were lysed in: 50mM Tris-HCl pH: 8.0; 150mM NaCl; 5mM EDTA; 10% Glycerol; 50mM NaF; 0.1mM Na3VO4; 0.2% NP40 and stored at $-80°$C. The debris was pelleted and then the supernatant was collected. The pellet was extracted according to [159]. Briefly, the pellet was resuspended in 200 ul of 0.1 M NaOH, 0.05 M ETDA, 2% SDS, and 2% beta-mercaptoethanol and incubated at 90°C for 10 minutes. Acetic acid was added to 0.1M and vortexed followed by an additional incubation at 90°C for 10 minutes before clarification by centrifugation and Methanol/chloroform extraction. The pellet was resuspended in 100 mM Tris containing 0.1% sodium deoxycholate with TCEP at 5 mM. Free thiols were capped with n-ethylmaleimide. Excess reagent was removed by ultrafiltration with amicon-4 10 kDa centrifugal devices. The protein was then quantified and exchanged into 6M guanidine for digestion overnight by αLP. The digests were quenched by the addition of formic acid to 1%, followed by desalting by sep-pak (Waters, Milford, MA). Peptides were then fractionated with Electrostatic Repulsion-Hydrophilic Interaction Chromatography [160]. Fractions were assayed for protein concentration using a BCA assay and pooled into 18 fractions of equal protein concentration, evaporated to dryness and resuspended in 100 uL of 0.2% FA. Nano liquid chromatography tandem mass spectrometry (nLC-MS/MS) was performed with a LTQ XL mass spectrometer equipped with ETD. 10 ul of each fraction ($\approx$ 1 ug) was injected onto a 12 cm $\times$ 75 um I.D.C18 column prepared in house and eluted in 0.2% FA with a gradient of 5% to 40% ACN over 60 min followed by wash and re-equilibration totaling 90 minutes of MS data per run. The flow was split about 1:500 to a flow rate of about 250 nL/min. A survey scan was followed by data dependent fragmentation of the 4 most abundant ions with both CID and ETD with supplemental activation. The maximum MS/MS ion accumulation time was set to 100 ms. Fragmented precursors were dynamically

excluded for 45 seconds with one repeat allowed.

6.3.3 Comparison of MS-GF+ and Mascot+Percolator

We compared the numbers of identified PSMs at 1% FDR for MS-GF+ and Mascot+Percolator (Table 6.5 for database search parameters). Mascot+Percolator (Mascot version 2.3.02 supporting Percolator) was used for the comparison because it represents the current state-of-the-art method for peptide identification. We also tested several other tools like InsPecT and OMSSA but do not report their results because they identified significantly fewer PSMs as compared to Mascot+Percolator.

Figure 6.7 (a) shows the benchmarking results for the five human datasets generated with varying fragmentations and instruments [156], corresponding to the following spectral paths: (CID,Low,Standard,Trypsin), (CID,High,Standard,Trypsin), (ETD,Low,Standard,Trypsin), (ETD,High,Standard,Trypsin), and (HCD ,High,Standard,Trypsin). While Percolator greatly increases the number of identifications as compared to Mascot, for all these datasets, MS-GF+ identified significantly more PSMs (8-38%) than Mascot+Percolator. We also compared the number of identifications reported by the original study [156] which also used Mascot+Percolator along with in-house pre- and post-processing tool. In this comparison, MS-GF+ also showed an improved performance (identifying 16-55% more PSMs).

To figure out how each tool benefits from high-precision product ion peaks, for the 3 out of 5 human datasets representing HH spectra, we ran MS-GF+, Mascot+Percolator, and Mascot using the parameters for HL spectra, i.e., using 0.6 Da fragment mass tolerance for Mascot and Mascot+Percolator, and using the scoring model for low-precision spectra for MS-GF+. For every tool, the number of identifications was higher when the parameters for HH spectra were used, but the difference varied depending on the dataset and the tool (Figure 6.7 (b)). For example, the difference was the largest for (CID,High,Standard,Trypsin) for MS-GF+ but for (HCD,High,Standard,Trypsin) for Mascot and Mascot+Percolator.

Figure 6.7 (c) shows the comparison for the ten yeast datasets generated

with varying fragmentations (CID or ETD) and enzymes (Trypsin, LysC, ArgC, GluC, or AspN) [143]. Again, for all these datasets, MS-GF+ identified significantly more PSMs (34-168%) than Mascot+Percolator (Figure 6.7 (c)). In [143], using OMSSA (and in-house tools for pre- and post-processing), the authors reported for each dataset the number of identified peptides at 1% peptide-level FDR that are matched to proteins identified at 1% protein-level FDR. We compared these numbers with the numbers of identified peptides at 1% peptide-level FDR using MS-GF+ (Figure 6.7 (d)). Note that this comparison is unfair because peptide identifications by MS-GF+ were not filtered out according to the protein that they are matched to. However, even after considering that, the results show that for most of the datasets, MS-GF+ identified many more peptides than the original report.

To see whether our scoring model can capture the fragmentation propensities specific to phosphopeptides, we generated a scoring parameter set for (CID, Low,Phosphorylation,Trypsin) and compared the numbers of identified PSMs for MS-GF+ with and without using the phosphorylation-specific parameter set, Mascot+Percolator, and InsPecT [25, 89] equipping with a dedicated scoring model for (CID,Low,Phosphorylation,Trypsin) (Figure 6.8). Interestingly, without phosphorylation-specific scoring parameters, MS-GF+ outperformed both tools, identifying 37% and 44% more PSMs than Mascot+Percolator and InsPecT, respectively. With phosphorylation-specific parameters, MS-GF+ identified 9% more PSMs (and 12% more PSMs of phosphopeptides), confirming that our scoring model successfully captures phosphorylation-specific fragmentation propensities. This function to easily train modification-specific scoring parameters (or any other experimental protocol that changes the fragmentation propensities) will greatly benefit MS researchers studying protein post-translational modifications.

6.3.4 Using MS-GF+ to identify peptides produced by αLP

MS-GF+ was applied to the study of αLP using two *S. pombe* datasets corresponding to (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP). Prior to this study, the cleavage specificity of αLP was unknown. We ran Mascot+Percolator,

OMSSA, and MS-GF+ by specifying 'None' as an enzyme. Mascot+Percolator and OMSSA performed very poorly for this novel spectral type. For example, Mascot+Percolator identified only 871 PSMs for (CID,Low,Standard,αLP) and *no* PSM for (ETD,Low,Standard,αLP) at 1% FDR. In contrast, MS-GF+ identified 3,535 and 2,829 PSMs from the (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP) dataset using the scoring parameters for (CID,Low,Standard,Trypsin) and (ETD,Low,Standard,Trypsin), respectively (Figure 6.9). The poor performance of Mascot+Percolator is because the scoring functions of Mascot and OMSSA are not adequate (correct peptide did not attain the maximal score) for most of the spectra due to the large search space (i.e. no enzyme is specified and the precursor mass tolerance was 2.5Da). In fact, for the human dataset corresponding to (ETD,Low,Standard,Trypsin), when no enzyme was specified and precursor mass tolerance was 2.5Da, Mascot identified no PSM at 1% FDR, whereas MS-GF+ identified 10,937 PSMs, only 34% less as compared to the fully-tryptic search with 7 ppm precursor mass tolerance.

Using the identified PSMs by MS-GF+, we trained scoring parameters for (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP). When these αLP-specific scoring parameters were used, the number of identified PSMs further increased to 4,788 (+35%) and 3,313 (+17%) for (CID,Low,Phosphorylation,αLP) and (ETD, Low,Phosphorylation,αLP), respectively, showing the usefulness of MS-GF+ for studies of new proteases.

αLP represents a new protease alternative to trypsin, greatly increasing the PTM and protein sequence coverages, but generating spectra with unusual fragmentation propensities. We emphasize that the capabilities of αLP are not obvious when Mascot+Percolator or another tool is used, because they fail to identify αLP peptides. The details on αLP protease will be discussed in a separate paper.

6.3.5 Running time of MS-GF+

We measured the running time of MS-GF+ and Mascot+Percolator for LL, HL, and HH spectra using the human datasets corresponding to (CID,Low,Stan-

dard,Trypsin) and (HCD,High,Standard,Trypsin), separately for a fully-tryptic search and a semi-tryptic search. For all the searches, MS-GF+ showed similar running times as compared to Mascot+Percolator (Figure 6.10).

6.3.6 Comparison of MS-GF+ with spectral library search

We compared the performance of MS-GF+ with the leading spectral library search tool SpectraST [6]. For this comparison, we downloaded the NIST spectral library (release date May 26th, 2011) containing 310,688 spectra generated from 190,539 unique peptides, and constructed a sequence database containing these 190,539 unique peptides. For the human dataset corresponding to (CID, Low,Standard,Trypsin), we ran MS-GF+ against this database by specifying the following set of common variable modifications: Oxidation of Met, Pyro-glu of Gln and Glu, Acetylation of protein N-term, Pyro-carbamidomethylation of Cys. Due to the reduced search space, MS-GF+ identified 19,002 PSMs in this search, 13% more as compared to the search against the entire IPI human database (Figure 6.11). For the same dataset, we also ran SpectraST search against the NIST spectral library and compared the number of identified PSMs with MS-GF+.[9] To get the best result, we ran SpectraST 5 times with varying precursor mass tolerance values (0.05Th, 0.1Th, 1Th, 2Th, and 3Th), and selected the maximum number of identifications obtained when 2 Th was used. Interestingly, SpectraST identified almost the same number PSMs (18,999 PSMs) as compared to MS-GF+. This indicates that if MS-GF+ use the *peptide* sequence information in the spectral library, it shows comparable performance without using any information on the *spectra* in the library.[10] We also compared the running time of SpectraST and MS-GF+ (against the database containing the library peptides), and both tools showed similar running times.

[9] Artificial decoy spectral library [161] was used to estimate the FDR.

[10] The search space was larger for MS-GF+ because it contains subsequences of peptides and all possible peptide variants with respect to the modifications.

6.4 Discussion

Our analysis showed that for diverse types of spectral datasets, MS-GF+ identifies more PSMs as compared to other existing tools for peptide identification including database search tools like Mascot, OMSSA, and InsPecT, a statistical modeling tool Percolator, and even a spectral library search tool SpectraST. In addition, MS-GF+ is as fast as other tools. Furthermore, MS-GF+ simplifies the computational pipeline for peptide identification because it does not require any additional pre- or post-processing tool.

The comparable performance of MS-GF+ over SpectraST indicates that access to previously identified spectra does not necessarily translate into significant improvement in accurate peptide identification. It implies that scoring methods that compare Spectrum-Spectrum Matches (SSMs) are also important. In contrast to the highly sophisticated methods for PSM scoring used in database searches, the library SSM scoring has not matured enough and is largely based on simple spectral cosine scores rather than statistical significance. We emphasize that the generating function approach for accurately computing E-values significantly contributes to the improved performance of MS-GF+. For example, when E-values instead of MS-GF scores were used to cut-off the results, the number of identified PSMs increased approximately by 70%, 50%, and 20% for LL, HL, and HH spectra, respectively. The performance of spectral library searching tools will thus be improved by developing rigorous methods for computing statistical significance of SSMs.

In a recent review, Noble and MacCoss pointed out that "the field (of MS) is still missing a generic analysis platform that can be adapted automatically and in a principled fashion to handle spectra produced by any given fragmentation protocol" [162]. With MS-GF+, we believe the field has made a step towards achieving this goal.

6.5 Acknowledgements

This chapter is in preparation for publication as "MS-GF+: Universal and Sensitive Database Search Tool for Mass Spectrometry". S. Kim, and Pavel Pevzner, in preparation. The dissertation author is the primary author of this paper.

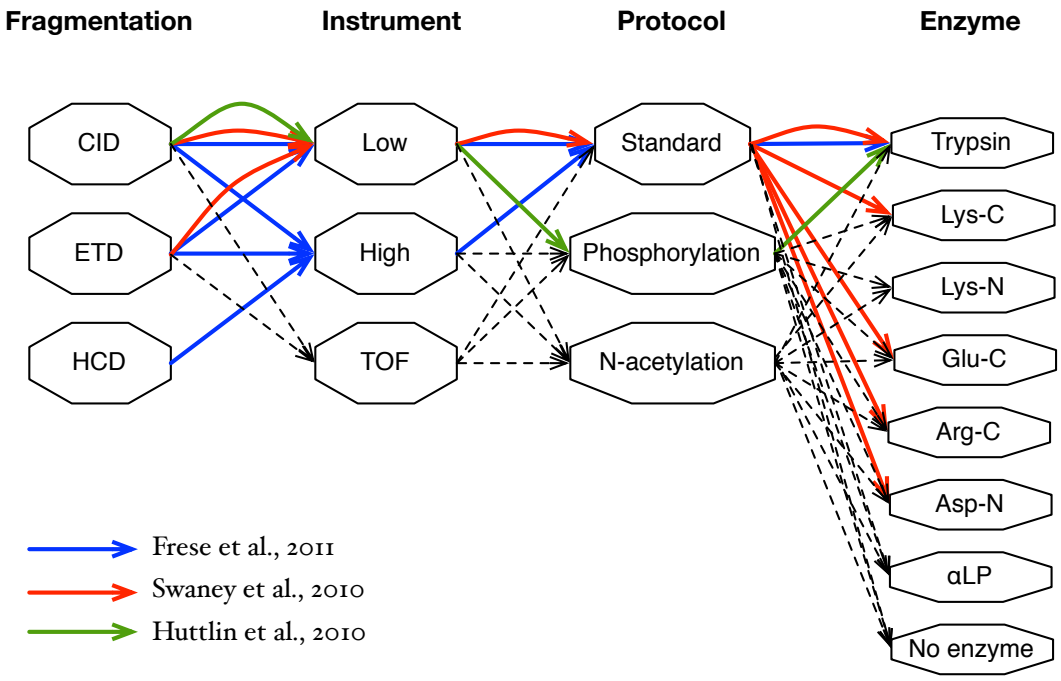
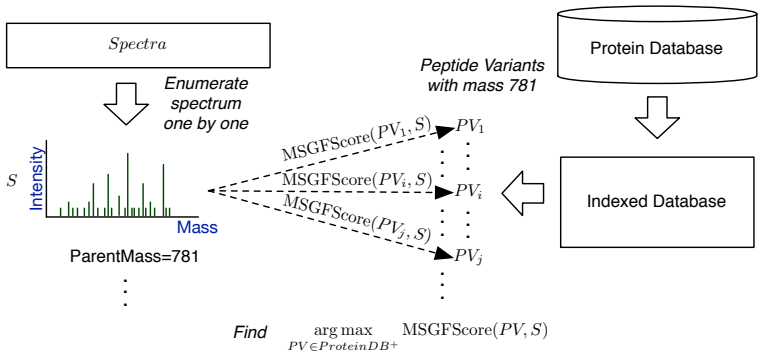
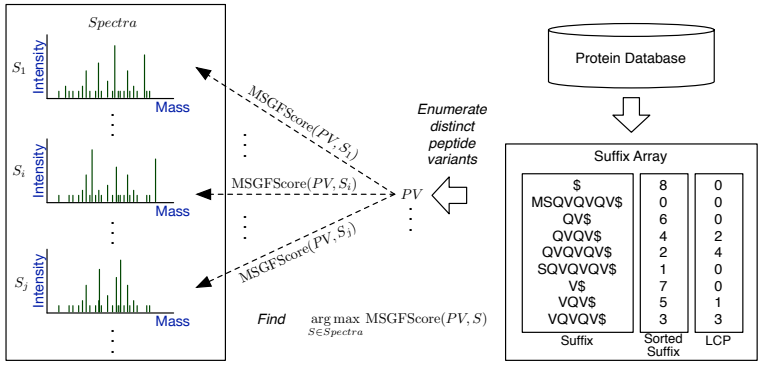


Figure 6.1: Spectral types as paths in the graph representing possible choices of the fragment method (Fragmentation), the instrument measuring product ion m/z (Instrument), the protocol used to prepare a sample (Protocol), and the enzyme used to digest proteins (Enzyme). 'Low' in Instrument indicates low-resolution instruments (e.g. linear ion-trap), 'High' indicates high-resolution instruments (e.g. Orbitrap, FT-ICR), and 'TOF' indicates time-of-flight instruments. 'Phosphorylation' and 'N-acetylation' in Protocol indicate that spectra are generated from phosphopeptides and peptides containing N-terminal acetylation, respectively. A path in the graph represents a spectral type. For example, the green path (CID,Low,Phosphorylation,Trypsin) represents low-precision CID spectra of trypsin digests generated from a sample enriched for phosphopeptides. The blue, red, and green paths represent spectral types of the spectral datasets used in recent studies by Frese et al. [156], Swaney et al. [143], and Huttlin et al. [157], respectively. Different combinations of analysis tools were used for different studies. Frese et al. used an in-house tool for peak filtering, de-isotoping, and charge deconvolution, Mascot for database search, Percolator for re-scoring, and RockerBox [163] for peptide-level FDR control. Swaney et al. used an in-house tool for peak filtering, OMSSA [13] for database search, and an in-house tool for for both peptide- and protein-level FDR control. Huttlin et al. used an in-house tool for re-calibrating peak masses, Sequest for database search, an in-house tool for re-scoring, and peptide- and protein-level FDR control. The same datasets were analyzed by MS-GF+ without using any additional tool using scoring parameters trained separately for different spectral types.



(a)



(b)

Figure 6.2: Two approaches for searching a protein database. (a) MS-GFDB's approach to use indices from peptide masses to peptides. (b) MS-GF+'s approach to use suffix arrays. See Online Methods for details.

x \ y	0	1
0	θ	$1-\rho$
1	$1-\theta$	ρ

(a)

x \ y,z	0,0	0,1	1,0	1,1
0	β_1	β_2	β_3	$1-\alpha$
1	$1-\beta_1$	$1-\beta_2$	$1-\beta_3$	α

(b)

Figure 6.3: (a) Probability $\text{Prob}_\text{V}(x|y)$ of a peptide character y generating a vertex label x. (b) Probability $\text{Prob}_\text{E}(x|y,z)$ of peptide characters y and z generating an edge label x.

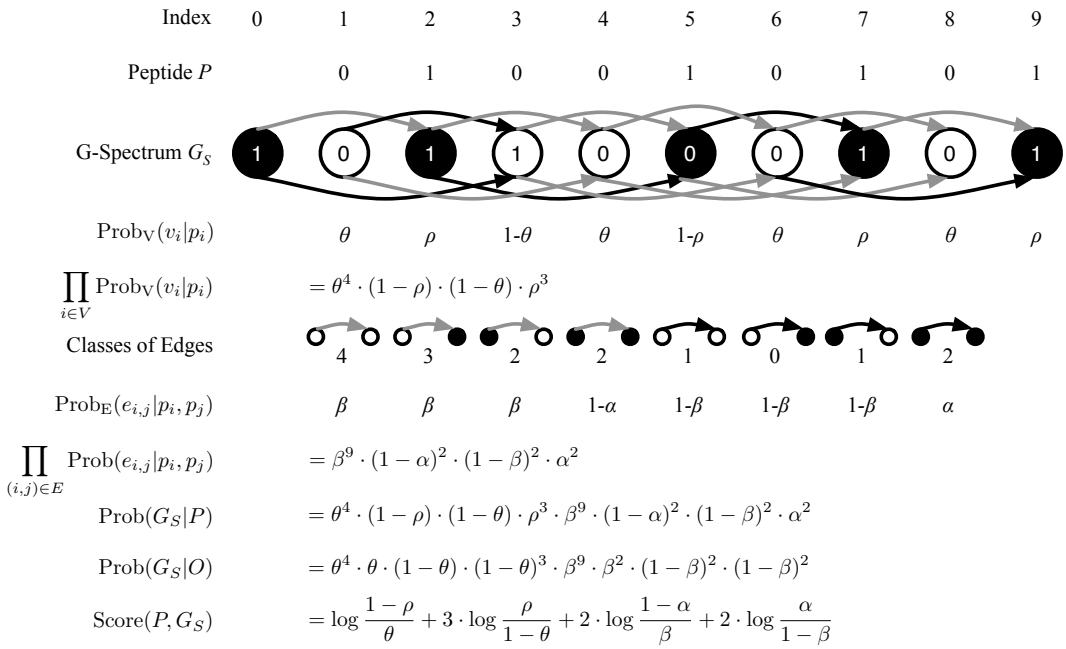


Figure 6.4: Illustration of the MS-GF+ scoring. A simplified amino acid model (only two amino acids A and B with masses 2 and 3, respectively) was used. The peptide ABAA is converted into its Boolean representation $P = 010010101$. The spectrum S is converted into a labeled DAG G_S. The number in the vertex represents its label. The color of the edge represents its label (0 for grey and 1 for black). The vertex i is colored depending on the peptide character i (white for 0 and black for 1). The procedure to compute $\text{Score}(P, G_S)$ is illustrated. All edges are partitioned into 8 classes depending on $e_{i,j}$, p_i, and p_j. For example, there are four edges with $e_{i,j} = p_i = p_j = 0$.

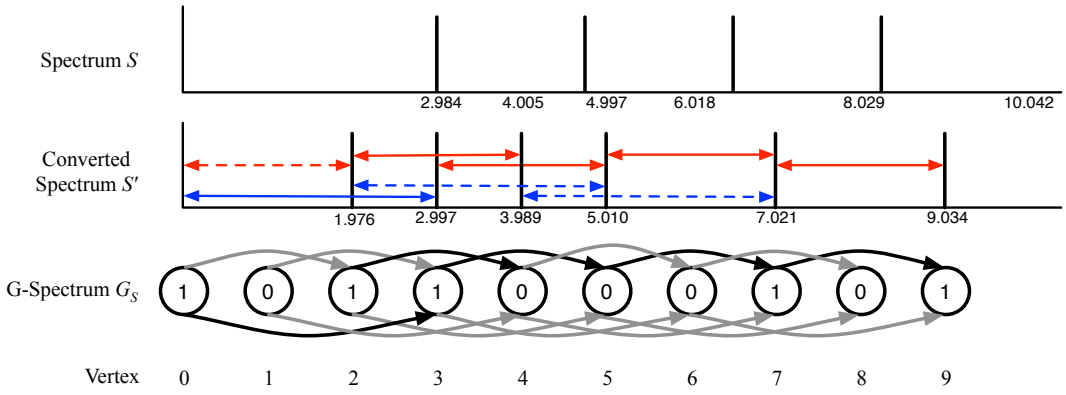
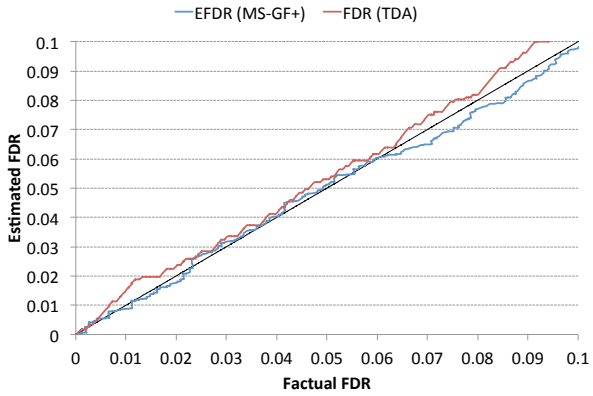
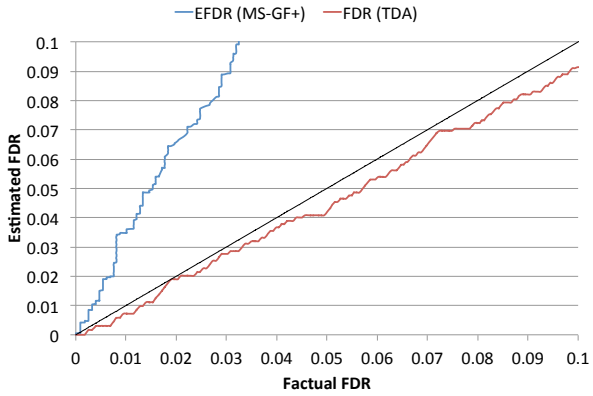


Figure 6.5: Illustration of the procedure constructing a G-spectrum. For simplification, a simplified amino acid model (only two amino acids with real masses 2.012 and 2.996) was used. Assume that only singly-charged b-ion with a real offset 1.008 contributes to the scoring. The spectrum S is given and converted into S' by shifting each peak by 1.008 to the left. Each arrowed line in S' represents a pair of peaks separated approximately by 2 Da (blue) or 3 Da (red) that form a duo (solid) or does not form a duo (dashed) for a fragment mass tolerance 0.01 Da. A G-spectrum G_S is constructed from S'. The number in the vertex represents its label. The color of the edge represents its label (0 for grey and 1 for black).



(a) LL spectra



(b) HL spectra

Figure 6.6: Accuracy of the EFDR reported by MS-GF+ and the FDR via TDA for LL spectra (a) and HL spectra (b). The factual FDR was used as an estimator of true FDR. The EFDR is accurate for LL spectra but biased for HL spectra (and also HH spectra).

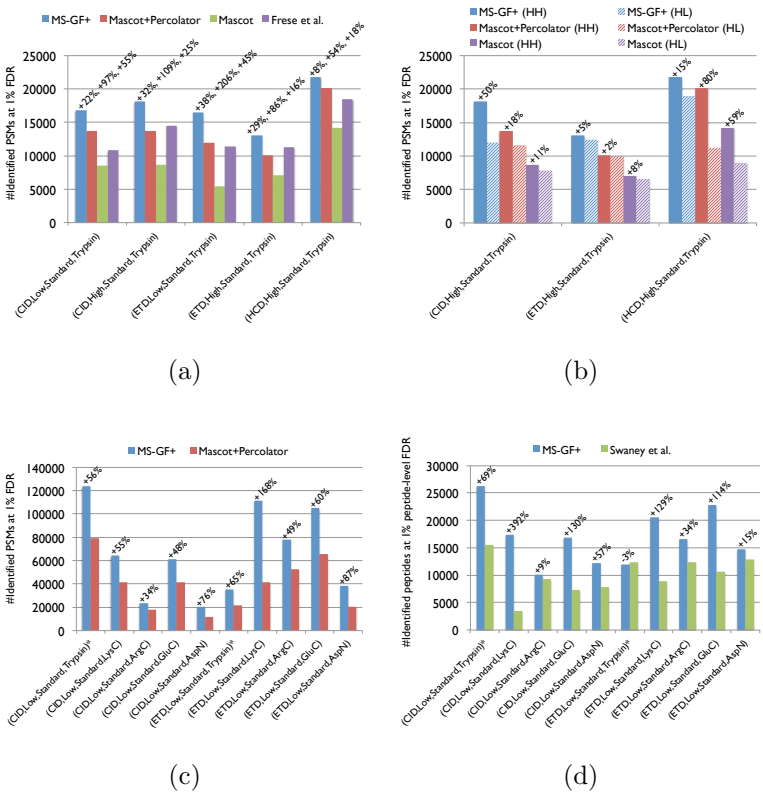


Figure 6.7: Comparison of MS-GF+ and other tools for diverse spectral types. The numbers of identified PSMs (a-c) or peptides (d) at 1% FDR are shown. Numbers above bars represent the percentages of increase in the number of identifications for MS-GF+ compared to other tools. (a) Results for the human datasets with varying fragmentations and instruments. MS-GF+, Mascot+Percolator, and Mascot results are shown along with the results in [156]. (b) Increase in the number of identifications due to the availability of high-precision product ion peaks. For the three human datasets representing HH spectra, MS-GF+, Mascot+Percolator, and Mascot were run using search parameters for HL spectra. The results of these searches (denoted by HL) are compared with the numbers of identifications for the regular searches using search parameters for HH spectra (denoted by HH). (c) Results for the yeast datasets with varying fragmentations and enzymes. MS-GF+ and Mascot+Percolator results are shown. (d) Comparison of MS-GF+ and the results in [143] that used OMSSA along with in-house post-processing tools for the yeast datasets. The numbers of (unique) peptides at the peptide-level 1% are shown. In [143], only the number of identified peptides matched to proteins identified at 1% protein-level FDR was counted while for MS-GF+, the number of identified peptides was counted regardless of their matched proteins.

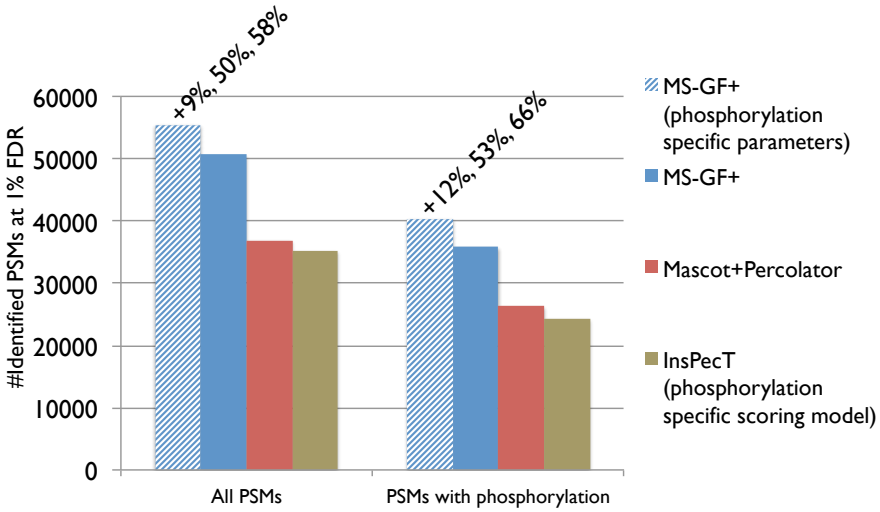


Figure 6.8: Comparison of MS-GF+, Mascot+Percolator, and InsPecT for the dataset corresponding to (CID,Low,Phosphorylation,Trypsin). The numbers of identified PSMs of all peptides (left) and phosphorylated peptides (right) at 1% FDR are shown. MS-GF+ was executed twice with the parameter set for (CID,Low,Standard,Trypsin) (denoted by MS-GF+) and for (CID,Low,Phosphorylation,Trypsin) (denoted by MS-GF+ (phosphorylation specific parameters)). Numbers above the bars are the percentages of increase in the number of identifications for MS-GF+ (phosphorylation specific parameters) as compared to MS-GF+, Mascot+Percolator, and InsPecT. MS-GF+ outperformed Mascot+Percolator and also InsPecT equipped with a phosphorylation-specific scoring model. When the scoring parameters for (CID,Low,Phosphorylation,Trypsin) was used, MS-GF+ identified 9% more PSMs (12% more PSMs of phosphopeptides) as compared to MS-GF+ using scoring parameters for (CID,Low,Standard,Trypsin).

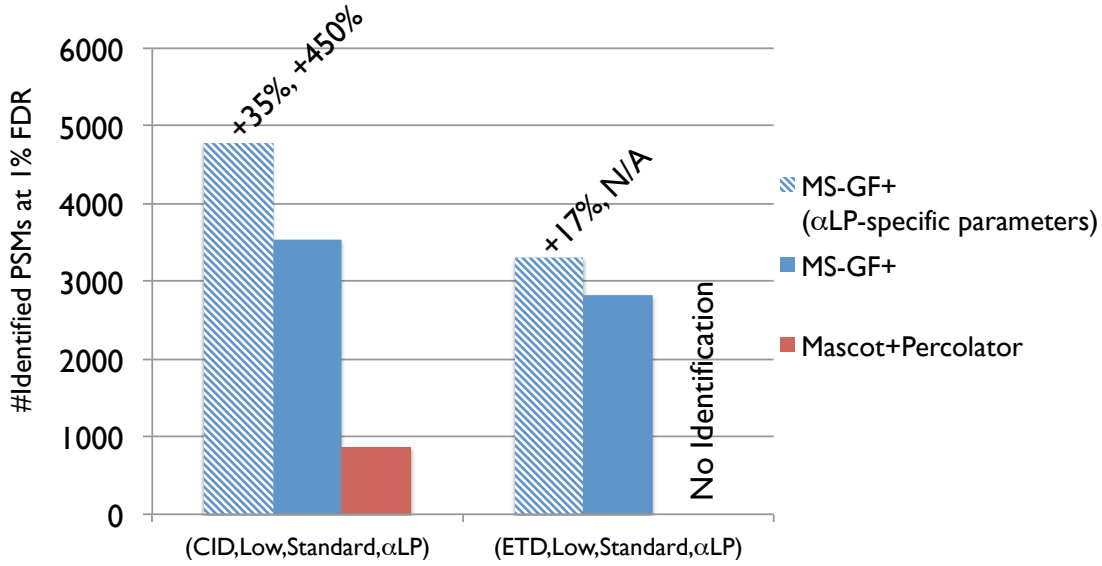
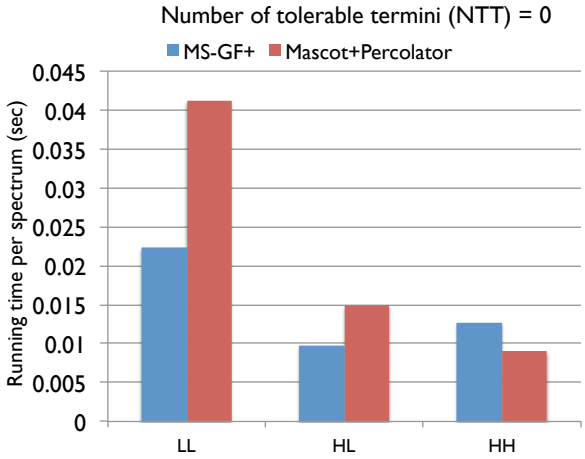
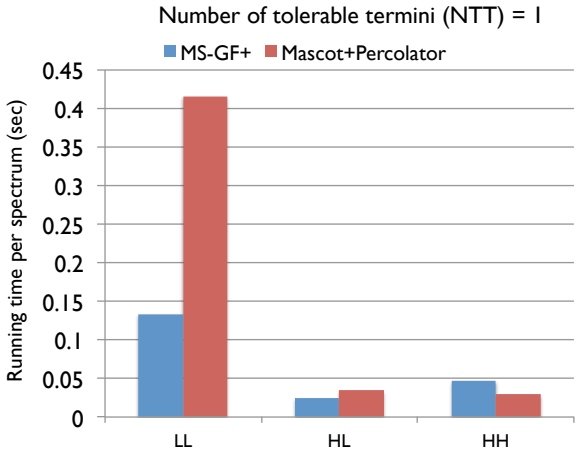


Figure 6.9: Comparison of MS-GF+, Mascot+Percolator for the datasets corresponding to (CID,Low,Phosphorylation,αLP) and (ETD,Low,Phosphorylation,αLP). MS-GF+ was executed twice with the parameter set for (CID,Low,Standard,Trypsin) (denoted by MS-GF+) and for (CID,Low,Standard,αLP) (denoted by MS-GF+ (αLP-specific parameters)). Shown on the top of the bars are the percentages of increase in the number of identified PSMs for MS-GF+ compared to Mascot+Percolator. Mascot+Percolator performed poorly on these datasets. In contrast, MS-GF+ identified a respectable number of PSMs.



(a)



(b)

Figure 6.10: Running time of MS-GF+ and Mascot+Percolator when NTT=0 (a) and NTT=1 (b). Average running time per spectrum is shown in second. LL, HL, HH represent LL, HL, and HH spectra, respectively. When NTT=0, MS-GF+ was 80% and 50% faster for LL and HL spectra, respectively, but 30% slower for HH spectra as compared to Mascot+Percolator. When NTT=1, MS-GF+ was 210% and 40% faster for LL and HL spectra, respectively, but 40% slower for HH spectra as compared to Mascot+Percolator.

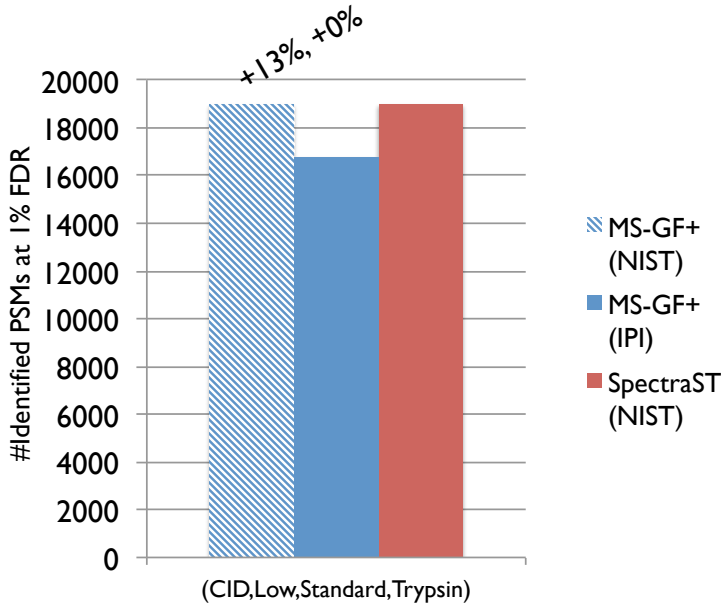


Figure 6.11: Comparison of MS-GF+ and SpectraST. For the MS-GF+ search, a database containing peptides in the NIST spectral library was constructed and searched (denoted by MS-GF+ (NIST)). For the SpectraST search, the NIST spectral library was searched (denoted by SpectraST (NIST library)). MS-GF+ results for the search against the IPI human database was shown for reference (denoted by MS-GF+ (IPI)). The numbers above bars represent the increase in the percentages of the number of identified PSMs for MS-GF+ (NIST) compared to MS-GF+ (IPI) and SpectraST (NIST library).

Table 6.1: Rescaling of amino acid masses. For each amino acid, shown are its nominal mass (NominalMass), real mass (RealMass), rescaled mass (RescaledMass), rounding error as the real mass minus nominal mass (ErrorRealMass), rounding error as the rescaled mass minus nominal mass (ErrorRescaledMass). Also the average of rounding errors and absolute rounding errors are shown. For every amino acid, rounding errors for rescaled masses are smaller (maximum 0.049 for Cys+57) as compared to real masses (maximum 0.101 for Arg). In addition, the average rounding error of 20 amino acids is only -0.004, meaning that even for long peptides the differences between their rescaled masses and nominal masses are usually small. Note that Carboxymethylated Cys (Cys+57) is considered instead of Cys here.

Residue	NominalMass	RealMass	RescaledMass	ErrorRealMass	ErrorRescaledMass
G	57	57.021	56.993	0.021	-0.007
A	71	71.037	71.002	0.037	0.002
S	87	87.032	86.989	0.032	-0.011
P	97	97.053	97.004	0.053	0.004
V	99	99.068	99.019	0.068	0.019
T	101	101.048	100.997	0.048	-0.003
L	113	113.084	113.028	0.084	0.028
I	113	113.084	113.028	0.084	0.028
N	114	114.043	113.986	0.043	-0.014
D	115	115.027	114.969	0.027	-0.031
Q	128	128.059	127.995	0.059	-0.005
K	128	128.095	128.031	0.095	0.031
E	129	129.043	128.978	0.043	-0.022
M	131	131.040	130.975	0.040	-0.025
H	137	137.059	136.990	0.059	-0.010
F	147	147.068	146.995	0.068	-0.005
R	156	156.101	156.023	0.101	0.023
C+57	160	160.031	159.951	0.031	-0.049
Y	163	163.063	162.982	0.063	-0.018
W	186	186.079	185.986	0.079	-0.014
			Average rounding error	0.057	-0.004
			Average absolute rounding error	0.057	0.017

Table 6.2: Database search parameters used for MS-GF+ and Mascot+Percolator searches. The number of tolerable termini (NTT) indicates the maximum number of peptide termini (0, 1, or 2) that is not consistent with the specificity of the enzyme. For example, for trypsin, NTT=0 means that only fully-tryptic peptides are considered in the search. For αLP digests, NTT was set to 2 because the cleavage specificity of αLP was unknown. For other non-tryptic enzymes, NTT was set to 1 (instead of 0 for trypsin) because they often produce peptides that are not perfectly consistent with their cleavage specificities. The number of ^{13}C was set to 1 to correct the error of choosing ^{13}C rather than ^{12}C peak during the MS1 peak detection. MS-GF+ does not take the fragment mass tolerance as an input, and rather implicitly assumes 0.5 Da tolerance for all spectra. For HH spectra, it benefits from accurate product ion peaks using the PRM graph scoring model (Online Method). MS-GF+ does not take the number of missed cleavages as an input, and rather allows peptides with any number of missed cleavages. For all the tools, the decoy search option was enabled for all searches to estimate the FDR. Note that a large precursor mass tolerance (50 ppm) was used for Mascot to provide sufficient training data for the Percolator algorithm [156]. Since Percolator gives a penalty to peptide identifications whose parent masses deviate from the precursor ion masses of the spectra, using a wide precursor mass tolerance increases rather than decreases the number of identifications. For Mascot searches for the human datasets, the precursor mass tolerance was set to 7 ppm because it produces more identifications. For the InsPecT search for the mouse dataset, the precursor and fragment mass tolerances were set to 2.5 Da and 0.5 Da, respectively, because using a narrower precursor mass tolerance decreased the number of identifications. To calculate the FDR, the spectral E-value, ion score, and F-score were used for MS-GF+, Mascot, InsPecT, respectively. For Percolator, instead of calculating the FDR, we used the q-value reported by Percolator.

Parameter	MS-GF+	Mascot+Percolator
Precursor mass tolerance	7 ppm	50 ppm
Precursor mass tolerance (*S. pombe*)	2.5Da	2.5Da
Fragment mass tolerance	N/A	0.6 Da (Low), 0.05 Da (High)
Number of missed cleavages	N/A	2
Fixed modification	Carbamidomethylation of Cys	
Variable modifications (Default)	Oxidation of Met, Acety of Prot N-term	
Variable modifications (Mouse)	Default + Phosph of Ser, Thr, and Tyr	
Variable modifications (*S. pombe*)	Default + Pyro-glu of Gln and Glu	
Number of tolerable termini (NTT)	0 (trypsin), 2 (αLP), 1 (others)	
Number of ^{13}C	1	
Protein Database (Human)	IPI human (ver. 3.87) + contaminants	
Protein Database (Yeast)	UniProt yeast (release 2012_02)	
Protein Database (Mouse)	IPI mouse (ver. 3.87) + contaminants	
Protein Database (*S. pombe*)	Uniprot *S. pombe*	

Bibliography

[1] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *J Proteomics*, vol. 73, pp. 2092–123, Oct 2010.

[2] R. Pathria, *Statistical Mechanics, 2nd edition*. Butterworth-Heinemann, Oxford, 1996.

[3] S. Kim, N. Gupta, N. Bandeira, and P. Pevzner, "Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra," *Mol. Cell. Proteomics*, vol. 8, pp. 53–69, Jan 2009.

[4] N. Castellana and V. Bafna, "Proteogenomics to discover the full coding content of genomes: a computational perspective," *J Proteomics*, vol. 73, pp. 2124–35, Oct 2010.

[5] D. Perkins, D. Pappin, D. Creasy, and J. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–67, Jan 1999.

[6] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold, "Development and validation of a spectral library searching method for peptide identification from ms/ms," *Proteomics*, vol. 7, pp. 655–67, Mar 2007.

[7] H. Lam, "Building and searching tandem mass spectral libraries for peptide identification," *Molecular & cellular proteomics : MCP*, vol. 10, p. R111.008565, Dec 2011.

[8] A. I. Nesvizhskii, O. Vitek, and R. Aebersold, "Analysis and validation of proteomic data generated by tandem mass spectrometry," *Nat Methods*, vol. 4, pp. 787–97, Oct 2007.

[9] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, A. Nesvizhskii, W. G. on Publication Guidelines for Peptide, and P. I. Data, "The need for guidelines in publication of peptide and protein identification data:

Working group on publication guidelines for peptide and protein identification data," *Molecular & cellular proteomics : MCP*, vol. 3, pp. 531–3, Jun 2004.

[10] R. A. Bradshaw, A. L. Burlingame, S. Carr, and R. Aebersold, "Reporting protein identification data: the next generation of guidelines," *Molecular & cellular proteomics : MCP*, vol. 5, pp. 787–8, May 2006.

[11] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search," *Anal. Chem.*, vol. 74, pp. 5383–92, Jan 2002.

[12] R. G. Sadygov and J. R. Yates, "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases," *Anal Chem*, vol. 75, pp. 3792–8, Aug 2003.

[13] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *J Proteome Res*, vol. 3, pp. 958–64, Jan 2004.

[14] S. Altschul, W. Gish, W. Miller, and E. Myers, "Basic local alignment search tool," *J. mol. Biol*, Jan 1990.

[15] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nat Methods*, vol. 4, pp. 207–14, Mar 2007.

[16] D. Fenyo, B. S. Phinney, and R. C. Beavis, "Determining the overall merit of protein identification data sets: rho-diagrams and rho-scores," *Journal of proteome research*, vol. 6, pp. 1997–2004, May 2007.

[17] R. Higdon, J. M. Hogan, G. V. Belle, and E. Kolker, "Randomized sequence databases for tandem mass spectrometry peptide and protein identification," *OMICS*, vol. 9, pp. 364–79, Jan 2006.

[18] R. E. Higgs, M. D. Knierman, A. B. Freeman, L. M. Gelbert, S. T. Patil, and J. E. Hale, "Estimating the statistical significance of peptide identifications from shotgun proteomics experiments," *Journal of proteome research*, vol. 6, pp. 1758–67, May 2007.

[19] S. A. Beausoleil, M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villén, J. Li, M. A. Cohn, L. C. Cantley, and S. P. Gygi, "Large-scale characterization of hela cell nuclear phosphoproteins," *Proc Natl Acad Sci USA*, vol. 101, pp. 12130–5, Aug 2004.

[20] W.-J. Qian, T. Liu, M. E. Monroe, E. F. Strittmatter, J. M. Jacobs, L. J. Kangas, K. Petritis, D. G. Camp, and R. D. Smith, "Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and sequest analysis: the human proteome," *Journal of proteome research*, vol. 4, pp. 53–62, Jan 2005.

[21] M. Waterman and M. Vingron, "Rapid and accurate estimates of statistical significance for sequence data base searches," *Proceedings of the National Academy of Sciences*, Jan 1994.

[22] D. Fenyö and R. C. Beavis, "A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes," *Anal Chem*, vol. 75, pp. 768–74, Feb 2003.

[23] J. Eriksson, B. T. Chait, and D. Fenyö, "A statistical basis for testing the significance of mass spectrometric protein identification results," *Anal Chem*, vol. 72, pp. 999–1005, Mar 2000.

[24] J. Eng, A. McCormack, and J. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976–89, Jan 1994.

[25] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, "Inspect: identification of posttranslationally modified peptides from tandem mass spectra," *Anal Chem*, vol. 77, pp. 4626–39, Jul 2005.

[26] N. Nagarajan, N. Jones, and U. Keich, "Computing the p-value of the information content from an alignment of multiple sequences," *Bioinformatics*, vol. 21 Suppl 1, pp. i311–8, Jun 2005.

[27] R. Graham, D. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, 1989.

[28] H. Wilf, *Generatingfunctionology*. Academic Press, Boston, MA, 1994.

[29] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner, "Identification of post-translational modifications by blind search of mass spectra," *Nat Biotechnol*, vol. 23, pp. 1562–7, Dec 2005.

[30] N. Bandeira, D. Tsur, A. Frank, and P. A. Pevzner, "Protein identification by spectral networks analysis," *Proc Natl Acad Sci USA*, vol. 104, pp. 6140–5, Apr 2007.

[31] J. Ng, N. Bandeira, W.-T. Liu, M. Ghassemian, T. L. Simmons, W. H. Gerwick, R. Linington, P. C. Dorrestein, and P. A. Pevzner, "Dereplication and de novo sequencing of nonribosomal peptides," *Nat Methods*, vol. 6, pp. 596–9, Aug 2009.

[32] J. Taylor and R. Johnson, "Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry," *Anal. Chem*, Jan 2001.

[33] V. Dancík, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner, "De novo peptide sequencing via tandem mass spectrometry," *J Comput Biol*, vol. 6, pp. 327–42, Jan 1999.

[34] T. Chen, M. Y. Kao, M. Tepel, J. Rush, and G. M. Church, "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry," *J Comput Biol*, vol. 8, pp. 325–37, Jan 2001.

[35] A. Frank and P. Pevzner, "Pepnovo: de novo peptide sequencing via probabilistic network modeling," *Anal. Chem*, Jan 2005.

[36] V. Bafna and N. Edwards, "On de novo interpretation of tandem mass spectra for peptide identification," *Proceedings of the seventh annual international conference on Research in computational molecular biology*, Jan 2003.

[37] B. Lu and T. Chen, "A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry," *J Comput Biol*, vol. 10, pp. 1–12, Jan 2003.

[38] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun Mass Spectrom*, vol. 17, pp. 2337–42, Jan 2003.

[39] M. Bern and D. Goldberg, "De novo analysis of peptide tandem mass spectra by spectral graph partitioning," *J Comput Biol*, vol. 13, pp. 364–78, Mar 2006.

[40] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann, "Novohmm: a hidden markov model for de novo peptide sequencing," *Anal Chem*, vol. 77, pp. 7265–73, Nov 2005.

[41] J. Grossmann, F. Roos, M. Cieliebak, and Z. Liptak, "Audens: a tool for automated peptide de novo sequencing," *J. Proteome Res*, Jan 2005.

[42] P. A. DiMaggio and C. A. Floudas, "De novo peptide identification via tandem mass spectrometry and integer linear optimization," *Anal Chem*, vol. 79, pp. 1433–46, Feb 2007.

[43] L. Mo, D. Dutta, Y. Wan, and T. Chen, "Msnovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry," *Anal Chem*, vol. 79, pp. 4870–8, Jul 2007.

[44] T. G. Dewey, "A sequence alignment algorithm with an arbitrary gap penalty function," *J Comput Biol*, vol. 8, pp. 177–90, Jan 2001.

[45] D. L. Tabb, C. G. Fernando, and M. C. Chambers, "Myrimatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis," *J Proteome Res*, vol. 6, pp. 654–61, Feb 2007.

[46] M. Bern, Y. Cai, and D. Goldberg, "Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry," *Anal. Chem.*, vol. 79, pp. 1393–400, Jan 2007.

[47] I. V. Shilov, S. L. Seymour, A. A. Patel, A. Loboda, W. H. Tang, S. P. Keating, C. L. Hunter, L. M. Nuwaysir, and D. A. Schaeffer, "The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra," *Mol Cell Proteomics*, vol. 6, pp. 1638–55, Sep 2007.

[48] N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith, and P. A. Pevzner, "Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation," *Genome Res*, vol. 17, pp. 1362–77, Sep 2007.

[49] R. Craig and R. C. Beavis, "Tandem: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, pp. 1466–7, Jun 2004.

[50] A. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem*, Jan 2003.

[51] D. L. Tabb, W. H. McDonald, and J. R. Yates, "Dtaselect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics," *Journal of proteome research*, vol. 1, pp. 21–6, Jan 2002.

[52] B. Zhang, M. C. Chambers, and D. L. Tabb, "Proteomic parsimony through bipartite graph analysis improves accuracy and transparency," *Journal of proteome research*, vol. 6, pp. 3549–57, Sep 2007.

[53] N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M. S. Lipton, M. Romine, V. Bafna, R. D. Smith, and P. A. Pevzner, "Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes," *Genome Res*, vol. 18, pp. 1133–42, Jul 2008.

[54] Y. Wan, A. Yang, and T. Chen, "Pephmm: a hidden markov model based scoring function for mass spectrometry database search," *Anal Chem*, vol. 78, pp. 432–7, Jan 2006.

[55] J. D. Venable and J. R. Yates, "Impact of ion trap tandem mass spectra variability on the identification of peptides," *Anal Chem*, vol. 76, pp. 2928–37, May 2004.

[56] G. Alves and Y.-K. Yu, "Robust accurate identification of peptides (raid): deciphering ms2 data using a structured library search with de novo based statistics," *Bioinformatics*, vol. 21, pp. 3726–32, Oct 2005.

[57] A. Frank, S. Tanner, V. Bafna, and P. Pevzner, "Peptide sequence tags for fast database search in mass-spectrometry," *J Proteome Res*, vol. 4, pp. 1287–95, Jan 2005.

[58] B. T. Hansen, S. W. Davey, A.-J. L. Ham, and D. C. Liebler, "P-mod: an algorithm and software to map modifications to peptide sequences using tandem ms data," *Journal of proteome research*, vol. 4, pp. 358–68, Jan 2005.

[59] B. C. Searle, S. Dasari, M. Turner, A. P. Reddy, D. Choi, P. A. Wilmarth, A. L. McCormack, L. L. David, and S. R. Nagalla, "High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for ms/ms de novo sequencing results," *Anal Chem*, vol. 76, pp. 2220–30, Apr 2004.

[60] M. Mann and M. Wilm, "Error-tolerant identification of peptides in sequence databases by peptide sequence tags," *Anal Chem*, vol. 66, pp. 4390–9, Dec 1994.

[61] C. Liu, B. Yan, Y. Song, Y. Xu, and L. Cai, "Peptide sequence tag-based blind identification of post-translational modifications with point process model," *Bioinformatics*, vol. 22, pp. e307–13, Jul 2006.

[62] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Commun Mass Spectrom*, vol. 11, pp. 1067–75, Jan 1997.

[63] J. D. Jaffe, H. C. Berg, and G. M. Church, "Proteogenomic mapping as a complementary method to perform genome annotation," *Proteomics*, vol. 4, pp. 59–77, Jan 2004.

[64] D. E. Kalume, S. Peri, R. Reddy, J. Zhong, M. Okulate, N. Kumar, and A. Pandey, "Genome annotation of anopheles gambiae using mass spectrometry-derived data," *BMC Genomics*, vol. 6, p. 128, Jan 2005.

[65] R. Wang, J. T. Prince, and E. M. Marcotte, "Mass spectrometry of the m. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias," *Genome Res*, vol. 15, pp. 1118–26, Aug 2005.

[66] D. Fermin, B. B. Allen, T. W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G. S. Omenn, and D. J. States, "Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics," *Genome Biol*, vol. 7, p. R35, Jan 2006.

[67] A. Savidor, R. Donahoo, O. Hurtado-Gonzales, N. VerBerkmoes, M. Shah, K. Lamour, and W. McDonald, "Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation.," *J. Proteome Res.*, vol. 5, pp. 3048–3058, 2006.

[68] S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigó, S. P. Briggs, and V. Bafna, "Improving gene annotation using peptide mass spectrometry," *Genome Res*, vol. 17, pp. 231–9, Feb 2007.

[69] A. Siepel, M. Diekhans, B. Brejová, L. Langton, M. Stevens, C. L. G. Comstock, C. Davis, B. Ewing, S. Oommen, C. Lau, H.-C. Yu, J. Li, B. A. Roe, P. Green, D. S. Gerhard, G. Temple, D. Haussler, and M. R. Brent, "Targeted discovery of novel human exons by comparative genomics," *Genome Res*, vol. 17, pp. 1763–73, Dec 2007.

[70] A. Stark, M. Lin, P. Kheradpour, J. Pedersen, L. Parts, J. Carlson, M. Crosby, M. Rasmussen, S. Roy, A. Deoras, J. Ruby, J. Brennecke, E. Hodges, A. Hinrichs, A. Caspi, B. Paten, S. Park, M. Han, M. Maeder, B. Polansky, B. Robson, S. Aerts, J. van Helden, B. Hassan, D. Gilbert, D. Eastman, M. Rice, M. Weir, M. Hahn, Y. Park, C. Dewey, L. Pachter, W. Kent, D. Haussler, E. Lai, D. Bartel, G. Hannon, T. Kaufman, M. Eisen, A. Clark, D. Smith, S. Celniker, W. Gelbart, and M. Kellis, "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures," *Nature*, vol. 450, pp. 219–232, 2007.

[71] J. Ng and P. A. Pevzner, "Algorithm for identification of fusion proteins via mass spectrometry," *J Proteome Res*, vol. 7, pp. 89–95, Jan 2008.

[72] K. Chong, K. Ning, H. Leong, and P. Pevzner, "Modeling and characterization of multi-charge mass spectra for Peptide sequencing.," *J. Bioinform Comput. Biol.*, vol. 4, pp. 1329–1352, 2006.

[73] K. R. Clauser, P. Baker, and A. L. Burlingame, "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching," *Anal Chem*, vol. 71, pp. 2871–82, Jul 1999.

[74] X. Cao and A. I. Nesvizhskii, "Improved sequence tag generation method for peptide identification in tandem mass spectrometry," *J Proteome Res*, vol. 7, pp. 4422–34, Oct 2008.

[75] D. Gusfield, *Algorithms on Strings, Trees and Sequences: computer science and computational biology*. Cambridge University Press, 1997.

[76] M. Havilio, Y. Haddad, and Z. Smilansky, "Intensity-based statistical scorer for tandem mass spectrometry," *Anal. Chem*, Jan 2003.

[77] J. Colinge, A. Masselot, M. Giron, T. Dessingy, and J. Magnin, "Olav: towards high-throughput tandem mass spectrometry data identification," *Proteomics*, vol. 3, pp. 1454–63, Aug 2003.

[78] S. Kim, N. Gupta, and P. Pevzner, "Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases," *J. Proteome Res.*, vol. 7, pp. 3354–63, Jul 2008.

[79] N. Bandeira, J. V. Olsen, J. V. Mann, M. Mann, and P. A. Pevzner, "Multi-spectra peptide sequencing and its applications to multistage mass spectrometry," *Bioinformatics*, vol. 24, pp. i416–23, Jul 2008.

[80] C. Bartels, "Fast algorithm for peptide sequencing by mass spectroscopy," *Biological Mass Spectrometry*, Jan 1990.

[81] B. Lu and T. Chen, "A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications," *Bioinformatics*, vol. 19 Suppl 2, pp. ii113–21, Oct 2003.

[82] R. Matthiesen, M. B. Trelle, P. Højrup, J. Bunkenborg, and O. N. Jensen, "Vems 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins," *J Proteome Res*, vol. 4, pp. 2338–47, Jan 2005.

[83] B. Ma, K. Zhang, and C. Liang, "An effective algorithm for peptide de novo sequencing from ms/ms spectra," *J Comput Syst Sci*, vol. 70, pp. 418–430, Jan 2005.

[84] Z. Zhang, "De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation," *Anal Chem*, vol. 76, pp. 6374–83, Nov 2004.

[85] V. Bafna and N. Edwards, "Scope: a probabilistic model for scoring tandem mass spectra against a peptide database," *Bioinformatics*, vol. 17 Suppl 1, pp. S13–21, Jan 2001.

[86] A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev, and P. A. Pevzner, "De novo peptide sequencing and identification with precision mass spectrometry," *J Proteome Res*, vol. 6, pp. 114–23, Jan 2007.

[87] J. D. Jaffe, N. Stange-Thomann, C. Smith, D. DeCaprio, S. Fisher, J. Butler, S. Calvo, T. Elkins, M. G. FitzGerald, N. Hafez, C. D. Kodira, J. Major, S. Wang, J. Wilkinson, R. Nicol, C. Nusbaum, B. Birren, H. C. Berg, and G. M. Church, "The complete genome and proteome of mycoplasma mobile," *Genome Res*, vol. 14, pp. 1447–61, Aug 2004.

[88] N. J. Edwards, "Novel peptide identification from tandem mass spectra using ests and sequence database compression," *Mol Syst Biol*, vol. 3, p. 102, Jan 2007.

[89] S. H. Payne, M. Yau, M. B. Smolka, S. Tanner, H. Zhou, and V. Bafna, "Phosphorylation-specific ms/ms scoring for rapid and accurate phosphoproteome analysis," *J Proteome Res*, vol. 7, pp. 3373–81, Aug 2008.

[90] A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith, and P. A. Pevzner, "Clustering millions of tandem mass spectra," *J Proteome Res*, vol. 7, pp. 113–22, Jan 2008.

[91] A. Shevchenko, S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens, and K. G. Standing, "Charting the proteomes of organisms with unsequenced genomes by maldi-quadrupole time-of-flight mass spectrometry and blast homology searching," *Anal. Chem.*, vol. 73, no. 9, pp. 1917–1926, 2001.

[92] Y. Han, B. Ma, and K. Zhang;, "Spider: software for protein identification from sequence tags with de novo sequencing error," *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*, pp. 206 – 215, Jul 2004.

[93] V. J. Denef, M. B. Shah, N. C. Verberkmoes, R. L. Hettich, and J. F. Banfield, "Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis," *J Proteome Res*, vol. 6, pp. 3152–61, Aug 2007.

[94] A. M. Frank, "Predicting intensity ranks of peptide fragment ions," *J Proteome Res*, vol. 8, pp. 2226–40, May 2009.

[95] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proc Natl Acad Sci USA*, vol. 84, pp. 4355–8, Jul 1987.

[96] D. L. Tabb, A. Saraf, and J. R. Yates, "Gutentag: high-throughput sequence tagging via an empirically derived fragmentation model," *Anal Chem*, vol. 75, pp. 6415–21, Dec 2003.

[97] S. Sunyaev, A. Liska, A. Golod, and A. Shevchenko, "Multitag: multiple error-tolerant sequence tag search for the sequence-similarity identification of . . . ," *Anal. Chem*, Jan 2003.

[98] R. Durbin, S. R. Eddy, A. K., and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[99] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold, and D. B. Martin, "The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools," *J Proteome Res*, vol. 7, pp. 96–103, Jan 2008.

[100] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–402, Sep 1997.

[101] L. Xin, G. Lajoie, and B. Ma, "New method for the validation of de novo sequencing results," *ASMS 2008*, p. WP645, 2008.

[102] Zubarev, N. Kelleher, and F. McLafferty, "Electron capture dissociation of multiply charged protein cations. a nonergodic process," *J Am Chem Soc*, pp. 3265–66, Jan 1998.

[103] H. J. Cooper, K. Håkansson, and A. G. Marshall, "The role of electron capture dissociation in biomolecular analysis," *Mass spectrometry reviews*, vol. 24, pp. 201–22, Jan 2005.

[104] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," *Proc Natl Acad Sci USA*, vol. 101, pp. 9528–33, Jun 2004.

[105] S. D. Taverna, B. M. Ueberheide, Y. Liu, A. J. Tackett, R. L. Diaz, J. Shabanowitz, B. T. Chait, D. F. Hunt, and C. D. Allis, "Long-distance combinatorial linkage between methylation and acetylation on histone h3 n termini," *Proc Natl Acad Sci USA*, vol. 104, pp. 2086–91, Feb 2007.

[106] N. Khidekel, S. B. Ficarro, P. M. Clark, M. C. Bryan, D. L. Swaney, J. E. Rexach, Y. E. Sun, J. J. Coon, E. C. Peters, and L. C. Hsieh-Wilson, "Probing the dynamics of o-glcnac glycosylation in the brain using quantitative proteomics," *Nat Chem Biol*, vol. 3, pp. 339–48, Jun 2007.

[107] E. Appella and C. W. Anderson, "New prospects for proteomics–electron-capture (ecd) and electron-transfer dissociation (etd) fragmentation techniques and combined fractional diagonal chromatography (cofradic)," *FEBS J*, vol. 274, p. 6255, Dec 2007.

[108] H. Molina, D. M. Horn, N. Tang, S. Mathivanan, and A. Pandey, "Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry," *Proc Natl Acad Sci USA*, vol. 104, pp. 2199–204, Feb 2007.

[109] A. F. M. Altelaar, S. Mohammed, M. A. D. Brans, R. A. H. Adan, and A. J. R. Heck, "Improved identification of endogenous peptides from murine nervous tissue by multiplexed peptide extraction methods and multiplexed mass spectrometric analysis," *J Proteome Res*, vol. 8, pp. 870–6, Feb 2009.

[110] S. Mohammed, K. Lorenzen, R. Kerkhoven, B. van Breukelen, A. Vannini, P. Cramer, and A. J. R. Heck, "Multiplexed proteomics mapping of yeast rna polymerase ii and iii allows near-complete sequence coverage and reveals several novel phosphorylation sites," *Anal Chem*, vol. 80, pp. 3584–92, May 2008.

[111] R. G. Sadygov, D. M. Good, D. L. Swaney, and J. J. Coon, "A new probabilistic database search algorithm for etd spectra," *J Proteome Res*, vol. 8, pp. 3198–205, Jun 2009.

[112] N. Taouatas, M. M. Drugan, A. J. R. Heck, and S. Mohammed, "Straightforward ladder sequencing of peptides using a lys-n metalloendopeptidase," *Nat Methods*, vol. 5, pp. 405–7, May 2008.

[113] D. Eppstein, "Targeted scx based peptide fractionation for optimal sequencing by collision induced, and electron transfer dissociation," *J. Proteomics Bioinform.*, vol. 1, no. 8, pp. 379–88, 2008.

[114] R. A. Zubarev, A. R. Zubarev, and M. M. Savitski, "Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet?," *J Am Soc Mass Spectrom*, vol. 19, pp. 753–61, Jun 2008.

[115] D. L. Swaney, G. C. McAlister, and J. J. Coon, "Decision tree-driven tandem mass spectrometry for shotgun proteomics," *Nat Methods*, vol. 5, pp. 959–64, Nov 2008.

[116] M. L. Nielsen, M. M. Savitski, and R. A. Zubarev, "Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry," *Mol Cell Proteomics*, vol. 4, pp. 835–45, Jun 2005.

[117] M. M. Savitski, M. L. Nielsen, F. Kjeldsen, and R. A. Zubarev, "Proteomics-grade de novo sequencing approach," *J Proteome Res*, vol. 4, pp. 2348–54, Jan 2005.

[118] R. Datta and M. Bern, "Spectrum fusion: Using multiple mass spectra for de novo peptide sequencing," *LECTURE NOTES IN COMPUTER SCIENCE*, Jan 2008.

[119] A. Bertsch, A. Leinenbach, A. Pervukhin, M. Lubeck, R. Hartmer, C. Baessmann, Y. A. Elnakady, R. Müller, S. Böcker, C. G. Huber, and O. Kohlbacher, "De novo peptide sequencing by tandem ms using complementary cid and electron transfer dissociation," *Electrophoresis*, vol. 30, pp. 3736–47, Nov 2009.

[120] H. Molina, R. Matthiesen, K. Kandasamy, and A. Pandey, "Comprehensive comparison of collision induced dissociation and electron transfer dissociation," *Anal Chem*, vol. 80, pp. 4825–35, Jul 2008.

[121] D. M. Good, C. D. Wenger, G. C. McAlister, D. L. Bai, D. F. Hunt, and J. J. Coon, "Post-acquisition etd spectral processing for increased peptide identifications," *J Am Soc Mass Spectrom*, vol. 20, pp. 1435–40, Aug 2009.

[122] N. Taouatas, A. F. M. Altelaar, M. M. Drugan, A. O. Helbig, S. Mohammed, and A. J. R. Heck, "Strong cation exchange-based fractionation of lys-n-generated peptides facilitates the targeted analysis of post-translational modifications," *Mol Cell Proteomics*, vol. 8, pp. 190–200, Jan 2009.

[123] S. Gauci, A. O. Helbig, M. Slijper, J. Krijgsveld, A. J. R. Heck, and S. Mohammed, "Lys-n and trypsin cover complementary parts of the phosphoproteome in a refined scx-based approach," *Anal Chem*, vol. 81, pp. 4493–501, Jun 2009.

[124] J. Cox and M. Mann, "Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat Biotechnol*, vol. 26, pp. 1367–72, Dec 2008.

[125] Y. Huang, J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith, and V. H. Wysocki, "Statistical characterization of the charge state and residue dependence of low-energy cid peptide dissociation patterns," *Anal Chem*, vol. 77, pp. 5800–13, Sep 2005.

[126] M. M. Savitski, M. L. Nielsen, and R. A. Zubarev, "Side-chain losses in electron capture dissociation to improve peptide identification," *Anal Chem*, vol. 79, pp. 2296–302, Mar 2007.

[127] A. Keller, J. Eng, N. Zhang, X. jun Li, and R. Aebersold, "A uniform proteomics ms/ms analysis platform utilizing open xml file formats," *Mol Syst Biol*, vol. 1, p. 2005.0017, Jan 2005.

[128] K. Jeong, S. Kim, N. Bandeira, and P. Pevzner, "Gapped spectral dictionaries and their applications for database searches of tandem mass spectra," *Research in Computational ...*, Jan 2010.

[129] P. A. Pevzner, V. Dancík, and C. L. Tang, "Mutation-tolerant protein identification by mass spectrometry," *J Comput Biol*, vol. 7, pp. 777–87, Jan 2000.

[130] L. Käll, J. D. Storey, M. J. Maccoss, and W. S. Noble, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases," *J Proteome Res*, vol. 7, pp. 29–34, Jan 2008.

[131] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. Maccoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets," *Nat Methods*, vol. 4, pp. 923–5, Nov 2007.

[132] M. Brosch, L. Yu, T. Hubbard, and J. Choudhary, "Accurate and sensitive peptide identification with mascot percolator," *J Proteome Res*, vol. 8, pp. 3176–81, Jun 2009.

[133] D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold, and A. I. Nesvizhskii, "iprophet: Multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates," *Molecular & cellular proteomics : MCP*, Aug 2011.

[134] P. Ross, D. Pappin, A. Heck, and S. Mohammed, "Straightforward and de novo peptide sequencing by maldi-ms/ms using a lys-n ...," *Molecular & Cellular Proteomics*, Jan 2009.

[135] J. J. Coon, "Collisions or electrons? protein sequence analysis in the 21st century," *Anal Chem*, vol. 81, pp. 3208–15, May 2009.

[136] J. V. Olsen and M. Mann, "Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation," *Proc Natl Acad Sci USA*, vol. 101, pp. 13417–22, Sep 2004.

[137] P. J. Ulintz, B. Bodenmiller, P. C. Andrews, R. Aebersold, and A. I. Nesvizhskii, "Investigating ms2/ms3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence," *Mol Cell Proteomics*, vol. 7, pp. 71–87, Jan 2008.

[138] B. Domon, B. Bodenmiller, C. Carapito, Z. Hao, A. Huehmer, and R. Aebersold, "Electron transfer dissociation in conjunction with collision activation to investigate the drosophila melanogaster phosphoproteome," *J Proteome Res*, vol. 8, pp. 2633–9, Jun 2009.

[139] D. L. Swaney, C. D. Wenger, J. A. Thomson, and J. J. Coon, "Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry," *Proc Natl Acad Sci USA*, vol. 106, pp. 995–1000, Jan 2009.

[140] R. J. Chalkley, A. Thalhammer, R. Schoepfer, and A. L. Burlingame, "Identification of protein o-glcnacylation sites using electron transfer dissociation mass spectrometry on native peptides," *Proc Natl Acad Sci USA*, vol. 106, pp. 8894–9, Jun 2009.

[141] P. J. Boersema, S. Mohammed, and A. J. R. Heck, "Phosphopeptide fragmentation and analysis by mass spectrometry," *J Mass Spectrom*, vol. 44, pp. 861–78, Jun 2009.

[142] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, "Higher-energy c-trap dissociation for peptide modification analysis," *Nat Methods*, vol. 4, pp. 709–12, Sep 2007.

[143] D. L. Swaney, C. D. Wenger, and J. J. Coon, "Value of using multiple proteases for large-scale mass spectrometry-based proteomics," *Journal of proteome research*, vol. 9, pp. 1323–9, Mar 2010.

[144] S. M. M. Sweet, A. W. Jones, D. L. Cunningham, J. K. Heath, A. J. Creese, and H. J. Cooper, "Database search strategies for proteomic data sets generated by electron capture dissociation mass spectrometry," *J Proteome Res*, vol. 8, pp. 5475–84, Dec 2009.

[145] E. J. Hsieh, M. R. Hoopmann, B. Maclean, and M. J. Maccoss, "Comparison of database search strategies for high precursor mass accuracy ms/ms data," *J Proteome Res*, vol. 9, pp. 1138–43, Nov 2009.

[146] S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck, and P. A. Pevzner, "The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: Applications to database search," *Mol Cell Proteomics*, vol. 9, pp. 2840–52, Dec 2010.

[147] C. Zhou, H. Chi, L.-H. Wang, Y. Li, Y.-J. Wu, Y. Fu, R.-X. Sun, and S.-M. He, "Speeding up tandem mass spectrometry-based database searching by longest common prefix," *BMC Bioinformatics*, vol. 11, p. 577, Jan 2010.

[148] U. Manber and G. Myers, "Suffix arrays: a new method for on-line string searches," *SIAM J. on Computing*, vol. 22, pp. 935–48, Jan 1990.

[149] M. Bern, D. Goldberg, W. H. McDonald, and J. R. Yates, "Automatic quality assessment of peptide tandem mass spectra," *Bioinformatics*, vol. 20 Suppl 1, pp. i49–54, Aug 2004.

[150] S. Na and E. Paek, "Quality assessment of tandem mass spectra based on cumulative intensity normalization," *J Proteome Res*, vol. 5, pp. 3241–8, Dec 2006.

[151] K. Jeong, S. Kim, N. Bandeira, and P. A. Pevzner, "Gapped spectral dictionaries and their applications for database searches of tandem mass spectra," *Molecular & cellular proteomics : MCP*, vol. 10, p. M110.002220, Jun 2011.

[152] W. S. Noble, "How does multiple testing correction work?," *Nat Biotechnol*, vol. 27, pp. 1135–7, Dec 2009.

[153] N. Gupta, N. Bandeira, U. Keich, and P. A. Pevzner, "Target-decoy approach and false discovery rate: when things may go wrong," *J Am Soc Mass Spectrom*, vol. 22, pp. 1111–20, Jul 2011.

[154] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: A peptide search engine integrated into the maxquant environment," *Journal of proteome research*, vol. 10, pp. 1794–805, Feb 2011.

[155] A. A. Klammer, C. Y. Park, and W. S. Noble, "Statistical calibration of the sequest xcorr function," *J Proteome Res*, vol. 8, pp. 2106–13, Apr 2009.

[156] C. K. Frese, A. F. M. Altelaar, M. L. Hennrich, D. Nolting, M. Zeller, J. Griep-Raming, A. J. R. Heck, and S. Mohammed, "Improved peptide identification by targeted fragmentation using cid, hcd and etd on an ltq-orbitrap velos," *J Proteome Res*, vol. 10, pp. 2377–88, May 2011.

[157] E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villén, W. Haas, M. E. Sowa, and S. P. Gygi, "A tissue-specific atlas of mouse protein phosphorylation and expression," *Cell*, vol. 143, pp. 1174–89, Dec 2010.

[158] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, "Proteowizard: open source software for rapid proteomics tools development," *Bioinformatics*, vol. 24, pp. 2534–6, Nov 2008.

[159] T. von der Haar, "Optimized protein extraction for quantitative proteomics of yeasts," *PLoS ONE*, vol. 2, p. e1078, Jan 2007.

[160] P. Hao, T. Guo, X. Li, S. S. Adav, J. Yang, M. Wei, and S. K. Sze, "Novel application of electrostatic repulsion-hydrophilic interaction chromatography

(erlic) in shotgun proteomics: comprehensive profiling of rat kidney proteome," *J Proteome Res*, vol. 9, pp. 3520–6, Jul 2010.

[161] H. Lam, E. W. Deutsch, and R. Aebersold, "Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics," *Journal of proteome research*, vol. 9, pp. 605–10, Jan 2010.

[162] W. S. Noble and M. J. Maccoss, "Computational and statistical analysis of protein mass spectrometry data," *PLoS Comput Biol*, vol. 8, p. e1002296, Jan 2012.

[163] H. W. P. van den Toorn, J. Muñoz, S. Mohammed, R. Raijmakers, A. J. R. Heck, and B. van Breukelen, "Rockerbox: analysis and filtering of massive proteomics search results," *Journal of proteome research*, vol. 10, pp. 1420–4, Mar 2011.