UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational and Comparative Proteogenomics: Annotating Genomes and Proteomes using Tandem Mass Spectrometry

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

in

Bioinformatics and Systems Biology

by

Nitin Gupta

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Steven P. Briggs, Co-Chair
Professor Vineet Bafna
Professor Elizabeth E. Komives
Professor Milton H. Saier, Jr.

2009

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

Co-Chair

Chair

University of California, San Diego

2009

DEDICATION

To my parents and my dearest friend Parul

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables x

Acknowledgements xi

Vita and Publications xiv

Abstract of the Dissertation xvi

Chapter 1 Introduction 1
1.1 Improving gene annotations 2
1.2 Detecting Proteolytic cleavages 2
1.3 Identification of Post-translational modifications 5

Chapter 2 Statistical significance of protein identifications 7
2.1 Introduction 7
2.2 Methods 9
2.3 Results 10
2.3.1 Comparison of two-peptide and single-peptide rule . 10
2.3.2 Estimating statistical significance of protein identifications using spectral dictionaries 12
2.3.3 Accounting for the size of spectral dataset 16
2.3.4 Using FPR of protein identifications 20
2.4 Discussion 21

Chapter 3 Proteogenomics 32
3.1 Introduction 32
3.2 Results 33
3.2.1 Using MS/MS data to find expressed proteins 34
3.2.2 Using MS/MS datasets to improve gene annotations 35
3.2.3 Using MS/MS datasets to analyze mutations and modifications 37
3.3 Identification of Signal peptides and N-terminal methionine cleavages 39
3.3.1 Proteolytic events and non-covered peptides 41

3.3.2 Predicting N-terminal methionine cleavage 42
3.3.3 Predicting signal peptides 43
3.4 Conclusion 44

Chapter 4 Comparative Proteogenomics 52
4.1 Introduction 52
4.2 Results 53
4.2.1 Multiple Shewanella Genomes 53
4.2.2 Protein Identification 55
4.2.3 Resolving One-Hit-Wonders 55
4.2.4 Correcting Gene Predictions: Start Sites 58
4.2.5 Identification of programmed frameshifts and sequencing errors 59
4.2.6 Proteolytic events 63
4.2.7 Post-translational modifications 67
4.3 Methods 68
4.3.1 Peptide Identification 68
4.3.2 Analyzing late start codons 69
4.3.3 Correlated peptides 69
4.4 Identification of post-translational modifications 70
4.5 Discussion 71

Chapter 5 Does trypsin cut before proline? 72
5.1 Introduction 72
5.2 Methods 75
5.2.1 Doubly-Confirmed Cleavages 75
5.2.2 Post-digestion breakup 76
5.3 Results 76
5.4 Discussion 80

Chapter 6 Improving detection of proteolytic sites 82
6.1 Introduction 82
6.2 Results 83
6.2.1 Peptide Identification 83
6.2.2 Reliable cleavage sites 83
6.2.3 V8 protease specificity 85
6.2.4 CNBr specificity 86
6.2.5 Chymotrypsin specificity 86
6.2.6 Using MS-GeneratingFunction for identifying cleavage sites 87
6.2.7 Detection of putative in vivo regulatory proteolytic sites 88
6.3 Discussion 91

Chapter 7 Novel proteogenomic applications 98
7.1 Introduction 98
7.2 Identification of neuropeptides 98
7.2.1 Methods 99
7.2.2 Results 100
7.2.3 Discussion 105
7.3 Using high-resolution MS/MS data for detecting mutations and rare modifications 107
7.3.1 Methods 108
7.3.2 Mutations 109
7.3.3 Rare PTMs 112
7.4 Improving operon predictions using mass spectrometry . . . 114
7.4.1 Results 115
7.4.2 Discussion 117
7.5 Conclusion 118

References 119

LIST OF FIGURES

Figure 2.1: Identification of peptides in the *Shewanella* dataset using different approaches 23
Figure 2.2: Identification of proteins using different approaches 24
Figure 2.3: Protein identification using X!Tandem search results 25
Figure 2.4: Identification of proteins using the unique peptides only 26
Figure 2.5: Identification of proteins using MS-GF scores 27

Figure 3.1: The ribosomal protein L31 (SO_4120) 34
Figure 3.2: Alignment of identified peptides and the hypothetical TIGR protein SO_1175 36
Figure 3.3: Alignment of identified peptides with the intergenic region between TIGR proteins SO_2299 and SO_2300 45
Figure 3.4: List of common modifications 46
Figure 3.5: Selected PTMs supported either by studies in other bacterial genomes or by comparative genomics analysis 47
Figure 3.6: Position of D+46 modification 47
Figure 3.7: Peptides from the N-terminal portion of conserved hypothetical protein SO_3842 47
Figure 3.8: Distribution of the N-termini of all non-covered peptides 48
Figure 3.9: Fraction of peptides undergoing cleavage of N-terminal methionine 49
Figure 3.10: Sequence logo for the amino acid sequence motif of all signal peptides 50
Figure 3.11: Venn diagram of all signal peptide predictions 51
Figure 3.12: Sequence of phosphoribosylformylglycinamidine cyclo-ligase . . . 51

Figure 4.1: Expression of orthologous genes across the three species 54
Figure 4.2: Example of correlated one-hit-wonders in shared genes 56
Figure 4.3: Commonly observed configurations of peptides in alternative frame 61
Figure 4.4: Frameshift generated by sequencing error 62
Figure 4.5: An example of a programmed frameshift 63
Figure 4.6: A cleavage site located within a peptide ladder 64

Figure 5.1: Different types of doubly confirmed cleavages 77
Figure 5.2: Raw and normalized frequencies of 45 most frequent di-AA pairs . 79

Figure 6.1: Fraction of different amino acids at P1 position 92
Figure 6.2: Sequence logo for the observed cleavage sites 93
Figure 6.3: Sequence logo for the observed cleavages in V8 protease digests . . 94
Figure 6.4: Sequence logo for the observed cleavages in chymotrypsin digests . 95
Figure 6.5: The histogram of positions in the corresponding protein sequence of the proteolytic sites 96

Figure 7.1: Sequence logos for the N- and C- termini of the observed neuropeptides 101
Figure 7.2: Sequence logos after removing the sites explained with dibasic amino acids 102
Figure 7.3: Increased peptide and protein coverage by multiple digests 103
Figure 7.4: Peptides in prochromogranin A 104
Figure 7.5: Peptides in proenkephalin 106
Figure 7.6: Graphical illustration of a putative operon in an MS-Operon report . 116
Figure 7.7: Splitting an operon into two 116

LIST OF TABLES

Table 2.1: An intuitive description of some frequently-used terms 28
Table 2.2: *Compression* for different values of spectral probability threshold . 29
Table 2.3: Summary of spectral clusters 30
Table 2.4: Compression computed for the ten largest clusters 31
Table 2.5: Comparison between the observed and the expected number of protein identifications for different values of spectral probability threshold 31

Table 4.1: Protein identification results 55
Table 4.2: Expression signatures for shared genes. 56
Table 4.3: List of conserved proteolytic sites 66

Table 6.1: Frequency of different amino acids at P1 position in the double-confirmed cleavage sites 97

Table 7.1: Frequency of different types of cells in the frequency matrix 110
Table 7.2: Frequency of paired-mutation cells 111
Table 7.3: Number of ortholog pairs on which different PTMs were found to be conserved between *Shewanella* and *E. coli*. 113

ACKNOWLEDGEMENTS

First and foremost, I am grateful to Prof. Pavel Pevzner, my PhD advisor, who not only directed and supported my research projects, but also taught me how to think like a researcher. I have learnt numerous skills from him, some specific to bioinformatics and some more general, like writing, presenting and time-management, that will affect me in all walks of life. Being a brilliant and meticulous scientist on one hand and a compassionate and patient advisor on the other, he will be a role model for my future academic career.

I am thankful to Prof. Vineet Bafna who has helped me as a guide, collaborator and friend at various stages of my research. I also thank Prof. Steven Briggs, Prof. Elizabeth Komives and Prof. Milton Saier for their guidance in shaping my research and for serving on my committee. I have had the opportunity to collaborate with many excellent researchers, most importantly Prof. Richard Smith at Pacific Northwest National Labs, who generously made several mass spectrometry datasets generated in his lab available for my research projects. I am also grateful to my other collaborators from PNNL, including Margaret Romine, Mary Lipton, Kim Hixson, Navdeep Jaitly, Joshua Adkins and David Culley. This work would not have been possible without the help of my collaborators Vivian Hook, Steve Bark, Sangtae Kim, Steven Tanner, Nuno Bandeira, Jeff Smith, Qing Sun, Wenhong Zhu, Louis Wich, Doug Lu, Jesse Rodriguez, Jamal Benhamida, Vipul Bhargava, Ian Kerman, Liz Kain, Jian Wang, Daniel Goodman, Noah Ollikainen, Ngan Nguyen, Daniel O'Connor, Laurent Taupenot, Andrei Osterman, Robert Edwards and Stefano Bonissone. I also thank members of the Pevzner and Bafna laboratories and students in the Bioinformatics and Systems Biology program for being helpful and supportive colleagues. I am very grateful to Laura Gracia for her continuous support as the graduate coordinator for the program.

The support of my family and friends cannot be over-stated. My parents, Mr. Subhash Chand Gupta and Mrs. Shashi Gupta, have provided me unfailing support at every stage of my life. I am also indebted to my friends Himanshu Khatri, Aneesh Subramanian, Nikhil Karamchandani, Nikhil Rasiwasia, Ankit Srivastava, Yash Kapoor, Nitin Mangal, Punit Kishore, Devashish Fuloria, Nikhil Agrawal, Rajat Mittal, and many others. Last, I am more grateful to Parul – my closest friend, guide, inspiration and support

in last three years – than I can express in words. Thank you for always being there.

Chapter 2 is, in part, a reprint of the paper “False discovery rates of protein identifications: a strike against the two-peptide rule. N. Gupta and P.A. Pevzner (2009). Journal of Proteome Research. 8(9):4173-81”. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is, in part, a reprint of the paper “Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. 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 (2007). Genome Research. 17(9):1362-77”. The dissertation author was the primary investigator and first author of this paper jointly with Steven Tanner.

Chapter 4 is, in part, a reprint of the paper “Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes. 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 (2008). Genome Research. 18:1133-1142”. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is, in part, a reprint of the paper “Does trypsin cut before Proline? J. Rodriguez, N. Gupta, R.D. Smith and P.A. Pevzner (2008). Journal of Proteome Research. 7(1):300-5”. The dissertation author was the second author and investigator of this paper.

Chapter 6 is, in part, a reprint of the submitted paper “Analyzing protease specificity and detecting in vivo proteolytic events using tandem mass-spectrometry. N. Gupta, K. K. Hixson, D. E. Culley, R. D. Smith and P. A. Pevzner”. The dissertation author was the primary investigator and author of this research.

Chapter 7 includes parts from multiple manuscripts in preparation in which the disseration author is the first or the second author. These manuscripts include “Evaluation of Alternative Neuropeptide Processing in Human and Bovine Dense-Core Secretory Granules by Mass Spectrometry-Based Neuropeptidomics. N. Gupta, S. Bark, W.D. Lu, L. Taupenot, D. O’Connor, P.A. Pevzner, V. Hook”, “MS-Operon: Using tandem mass spectrometry for operon prediction and validation. L. Wich and N. Gupta” and “Discovery of mutations and rare modifications using mass spectrometry. N. Gupta, R. D. Smith and P. A. Pevzner”.

VITA

2004 B. Tech. in Computer Science and Engineering, Indian Institute of Technology, Kanpur, India

2009 Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego, USA

PUBLICATIONS

N. Gupta, R.D. Smith, P.A. Pevzner. Discovery of mutations and rare modifications using mass spectrometry. *In preparation*.

N. Gupta, S. Bark, W.D. Lu, L. Taupenot, D. O'Connor, P.A. Pevzner, V. Hook. Evaluation of Alternative Neuropeptide Processing in Human and Bovine Dense-Core Secretory Granules by Mass Spectrometry-Based Neuropeptidomics. *In preparation*.

S. Bonissone, N. Gupta and P.A. Pevzner. Comparative proteogenomics reveals a possible functional role of N-terminal Methionine Excision. *In preparation*.

L. Wich and N. Gupta. MS-Operon: Using tandem mass spectrometry for operon prediction and validation. *In preparation*.

N. Gupta, N. Bandeira and P.A. Pevzner. Target-decoy approach in proteomics: when things may go wrong. *Submitted*.

N. Gupta, K. K. Hixson, D. E. Culley, R. D. Smith and P. A. Pevzner. Analyzing protease specificity and detecting in vivo proteolytic events using tandem mass-spectrometry. *Submitted*.

N. Gupta and P.A. Pevzner (2009). False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*. 8(9):4173-81.

S. Kim, N. Gupta, N. Bandeira and P.A. Pevzner (2009). Spectral Dictionaries: Integrating De Novo Peptide Sequencing with Database Search of Tandem Mass Spectra. *Molecular and Cellular Proteomics*. 8(1):53-69.

B. Dost, T. Shlomi, N. Gupta, E. Ruppin, V. Bafna, and R. Sharan (2008). QNet: a tool for querying protein interaction networks. *Journal of Computational Biology*. 15(7):913-25.

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 (2008). Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes. *Genome Research*. 18:1133-1142.

S. Kim, N. Gupta and P.A. Pevzner (2008). Spectral Probabilities and Generating Functions of Tandem Mass Spectra: a New Approach to Peptide Identifications. *Journal of Proteome Research*. 7(8): 3354 - 3363.

J. Rodriguez, N. Gupta, R.D. Smith and P.A. Pevzner (2008). Does trypsin cut before Proline? *Journal of Proteome Research*. 7(1):300-5.

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 (2007). Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Research*. 17(9):1362-77.

K. Gaurav, N. Gupta and R. Sowdhamini (2005). FASSM: Enhanced Function Association in whole genome analysis using Sequence and Structural Motifs. *In Silico Biology*. 5:0040.

N. Gupta, N. Mangal and S. Biswas (2005). Evolution and similarity evaluation of protein structures in contact map space. *Proteins: Structure, Function and Bioinformatics*. 59(2):196-204.

A. Bhaduri, G. Pugalenthi, N. Gupta and R. Sowdhamini (2004). iMOT: an interactive package for the selection of spatially interacting motifs. *Nucleic Acids Research*. 32:W602-W605.

N. Gupta and A. Irback (2004). Coupled folding-binding versus docking: A lattice model study. *Journal of Chemical Physics*. 120: 3983-3989.

ABSTRACT OF THE DISSERTATION

Computational and Comparative Proteogenomics: Annotating Genomes and Proteomes using Tandem Mass Spectrometry

by

Nitin Gupta
Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2009

Professor Pavel A. Pevzner, Chair

Professor Steven P. Briggs, Co-Chair

Next generation sequencing techniques will soon lead to an explosive growth in the number of sequenced genomes and will turn manual gene and protein annotations into a luxury that can be afforded for only a small fraction of the newly sequenced genomes. In this work, we show that mass spectrometry can be a viable alternative high-throughput approach for accurate proteogenomic annotation. We present advances in statistical analysis of the reliability of peptide and protein identifications from mass spectrometry, and demonstrate how comparative analysis of multiple species or multiple experimental samples can provide additional insights for new biological discovery. We apply these approaches for correcting gene boundaries, discovering programmed frameshifts, N-terminal methionine cleavages, signal peptide cleavages and other in-vivo regulatory proteolytic events, post-translational modifications, analyzing specificity of individual proteases, detecting neuropeptides in brain tissue, and improving operon predictions. Based on this work, we recommend complementing genome sequencing projects by mass spectrometry-based proteogenomics.

Chapter 1

Introduction

While bacterial genome annotations have significantly improved in recent years, techniques for bacterial proteome annotation (including post-translational chemical modifications, signal peptides, other proteolytic events, etc.) are still in their infancy. In this work, we developed new approaches for analyzing mass spectrometry data to improve both genome and proteome annotations, focusing on computational and comparative *proteogenomics*. Computational, because we rely on improvements in computational analysis of data to derive new insights instead of requiring new experimental procedures. Comparative, because we make use of datasets from multiple organisms or multiple types of enzyme-digests to derive insights that are difficult to derive from individual samples.

In Gupta et al., 2009 [52], we presented advances in statistical analysis of protein error rates, showing that accurate false positive rates of *individual* proteins can be computed efficiently using the generating function approach [75, 74], compared to the false discovery rates computed using the traditional target-decoy approaches. We also provided evidence that the commonly accepted "two-peptide" rule negatively affects the number of protein identifications and therefore should be abandoned [52]. Below we provide an overview of some important proteogenomic applications (other applications are discussed in Chapter 7).

1.1 Improving gene annotations

Recent proliferation of low-cost DNA sequencing techniques will soon lead to an explosive growth in the number of sequenced genomes and will turn manual annotations into a luxury that can be afforded for only a small fraction of the newly sequenced genomes. The idea of querying a MS/MS dataset against a genome to identify protein coding genes has been used earlier in different settings [148, 78, 103, 42]. Bacterial genomes, with a simple gene structure, are a particularly attractive target for such methods. The identified peptides validate the predicted genes, correct erroneous gene annotations, and reveal some completely missed genes. Church and colleagues used proteomic data for genome analysis of relatively small bacterium, Mycoplasma pneumoniae [65], and later on the newly sequenced Mycoplasma mobile in which 26 genes were predicted exclusively based on proteomic data [64]. Similar efforts have been made for other bacterial genomes [140, 69]. Nevertheless, many significant technological challenges remain in using MS/MS for correcting gene annotation, discovering alternative start codons, detecting frameshifts, finding short genes etc. Detection and mapping of multiple peptides to a single gene can be evidence that it encodes a protein that is expressed. Similarly, multiple matches to a genomic region outside the boundary of genes can be used to detect new genes missed during genome annotation or to suggest that gene boundaries should be expanded. In Gupta et al., 2007 [53], we developed this simple idea to address the some of the questions raised above.

Previous proteogenomic approaches were limited to a single genome and did not take advantage of analyzing mass spectrometry data from multiple genomes at once. In Gupta et al., 2008 [51], we showed that such comparative proteogenomics approach (like comparative genomics) allows one to address the problems that remained beyond the reach of the traditional “single proteome” approach in mass-spectrometry.

1.2 Detecting Proteolytic cleavages

Proteolytic cleavage through cellular proteases is extremely important for many biological functions. While such cleavage is often specific and tightly regulated, protease activity in cells is relatively unexplored, primarily due to the lack of effective high-throughput technology to detect proteolytic events. Large MS/MS datasets offer an unprecedented opportunity to study in-vivo cleavage specificity by looking at over-represented non-tryptic peptides that may be manifestations of proteolytic events. In Gupta et al., 2007 [53], we analyzed such peptides to identify N-terminal methionine excisions and signal peptide cleavages.

Proteases are molecular scissors that play a critical role in the regulatory processes inside the cell as well as molecular tools in the laboratory. The defining characteristic of a protease is its specificity, i.e. the rule that determines the selection of its cleavage-substrates. Knowledge of specificity is important for understanding the function and mechanism of proteases and for their laboratory applications. One such application of proteases is in the form of digestive enzymes in mass spectrometry-based proteomics [1], where they are used to cleave proteins into smaller peptides that are easier to analyze than intact proteins. Trypsin is the most commonly used protease for this purpose, partly because of its well-defined and robust specificity rules [100]. As we argued in [112], having precise knowledge of specificity of the protease is important not only for peptide identification (many peptide identification tools incorporate specificity rules into their search algorithms), but is also critical in some emerging applications of mass spectrometry such as label-free analysis of regulatory proteolysis [131, 37, 51]. In such studies, the sample is digested with a protease with *known* specificity (e.g., trypsin, V8 protease, etc.) and a regulatory protease (e.g., a caspase) with the goal to discover the (*unknown*) specificity of the regulatory protease. Tandem mass spectrometry (MS/MS) is then employed to determine all cleavages in the resulting sample. Afterwards, one has to “subtract” expected in vitro cleavages (e.g., trypsin induced cleavages) from all found cleavages to identify the in vivo cleavages caused by the regulatory proteases. However, if the model of the protease specificity is even slightly inaccurate, these studies are likely to fail. For example, while the rule “trypsin cuts after R and K but not before P” is a reasonable description of trypsin specificity for most applications, it becomes inaccurate if one attempts to find in vivo proteolytic sites (since trypsin actually cuts before P albeit with reduced efficiency [112]). As a result, if one uses an inaccurate rule for trypsin specificity, the cuts before P will not be subtracted resulting in a surprising “discovery” of many in vivo cleavage sites before P. In reality, this “discovery” reveals limitations of the common rule describing trypsin specificity rather than a new protease activity.

Another area that requires detailed knowledge of protease specificity is the proteome wide analysis of in vivo proteolytic events in the sample subjected to trypsin with the goal to infer the natural proteolytic cleavages induced by various proteases (without attempting to infer the specificities of individual proteases). In the past, information about proteolysis has been mainly gained by performing in vitro experiments with individual proteins and proteases that may not represent true in vivo scenarios at the proteome-wide scale. Recently, Manes et al., 2007 [89], and Shen et al., 2008 [121] addressed the challenge of proteome-wide proteolysis analysis in the studies of native (short) peptides in *Saccharomyces cerevisiae* and *Salmonella enterica*. However, longer native peptides require digestion with trypsin or other proteases, and there is stil no software tool that can identify in vivo proteolytic sites from such digests.

Determining specificity of proteases has traditionally been a strenuous experimental process, and consequently, often limited to analysis of a small number of substrates [71]. Combinatorial library approaches address this shortcoming by employing large libraries of substrates treated by the protease [110, 54, 138], although analyzing the cleaved products from these libraries may require use of laborious fluorescence or sequencing technology. Mass spectrometry presents a rapid approach for sequencing a large number of substrates from a peptide library. Recently, Schilling and Overall, 2008 [119] described peptide libraries derived from human proteome that could be easily analyzed by mass spectrometry through standard database-search methods. This approach, however, required the use of biotin-labeling to separate the N-terminal and C-terminal side of the cleavage sites.

In Rodriguez et al. 2008 [112], we demonstrated that it is possible to determine accurate specificity rules for the enzyme used for digestion in a standard mass spectrometry experiment when analyzing large spectral datasets. This approach can be easily implemented (even if the data were generated for a different purpose), without the requirement of expensive labeling methods. In contrast to most proteomics approaches (that typically identify peptides with FDR 1% amd higher), this approach requires extremely accurate peptide identifications (typically FDR 0.1% and lower) since even a small fraction of incorrect assignments may contribute many (pseudo) cleavages that distort the analysis of protease specificity. Rodriguez et al. 2008 [112] used *doubly-confirmed* cuts to arrive at a reliable set of identified peptides. In Chapter 6, we further extended this approach to remove the doubly-confirmed requirement, making use of MS-GeneratingFunction [75]. In this work, we showed that comparative analysis of multiple digests allows one to reliably identify N-terminal methionine excisions, signal peptide cleavages and other putative proteolytic events, in the same way as we demonstrated in Gupta et al., 2008 [51] that comparative analysis of multiple species can be helpful in confirming proteolytic events.

1.3 Identification of Post-translational modifications

The current understanding of post-translational chemical modifications (PTMs) in bacteria is very limited even for well-studied organisms like *E. coli*. Any information that could be obtained about PTMs from large scale MS/MS studies can prove to be very important towards gaining an understanding of the molecular biology of bacterial genomes. In Gupta et al., 2007 [53], we analyzed the mass spectra using MS-alignment [135], and identified more than 10,000 modified peptides in *Shewanella*, spanning at least 25 well-covered modifications and many other less-frequent ones.

Diphthamide is an extremely rare histidine modification that appears on a single gene (translation elongation factor 2) in the entire human genome [93, 139, 84]. Diphthamide is a target of diphtheria toxin and its position is conserved over a billion years of evolution (from yeast to human). However, systematic identification of *new* important and rare modifications remains a difficult, if not impossible, problem in shotgun proteomics experiments. While algorithms for *blind* searches for unexpected modifications have been developed (e.g., MS-Alignment [136] and ModifiComb [117]), they had to rely on the “strength in numbers” principle to distinguish real modifications from computational artifacts. As a result, the biologically important modifications that appear only a few times in the genome are likely to be classified as computational artifacts. For example, each of 25 most common modifications in So appear on at least 39 sites in the genome [53] pushing rare modifications to the twilight zone of the statistical significance. In Gupta et al., 2008 [51], we showed that comparative proteogenomics allows one to identify putative rare modifications in shotgun proteomics experiments.

Chapter 2

Statistical significance of protein identifications

2.1 Introduction

Tandem mass spectrometry (MS/MS) database search tools are routinely used for peptide identifications [1]. The results often include many false positives [20], and a common approach for estimating the False Discovery Rate (FDR) of peptide identifications is based on the use of randomized decoy databases [36, 68]. The reported set of peptide identifications is determined by varying the score cutoff to achieve the desired FDR. Intuitively, a protein must be present in the sample if a peptide within it is identified (assuming the same peptide is not present elsewhere in the proteome). However, since peptide identification tools generate some false peptides, many researchers are cautious about the "one-hit-wonders" [102, 21, 17], making it a common practice to report only proteins with at least two peptides as reliable identifications (proteins with single peptide identifications are often ignored, or delegated to the Supplementary materials). The "two-peptide" rule has done the field a great service by providing a stringent criterion for reporting proteomic data. However, while the "two peptide" rule seems well-intentioned and reasonable, we are unaware of any *theoretical* studies supporting this rule. Indeed, the question whether two identified peptides with scores x and y (within the same protein) represent a better evidence for expression of this protein than a single peptide with score z (that is larger than x and y) depends on parameters x, y, z (and the protein length) and remains poorly addressed. Below we show that the "two-peptide" rule is inferior to the "single-peptide" rule that takes into account one-hit-wonders (with appropriately chosen score threshold z). We therefore argue that the "two-peptide" bias should be removed and that protein identifications based on single peptides should be treated *at par* with identifications based on multiple peptides, instead of salvaging them through post-processing [56].

Gupta et al., 2007 [53] estimated that 80% of the one-hit-wonders (proteins with a single identified peptide) in proteogenomics study of *Shewanella oneidensis* MR-1 are likely to be expressed. Comparative analysis of three *Shewanella* species revealed that many of these one-hit-wonders are actually observed as orthologs in multiple species, providing further support for their expression [51]. These observations indicate that the "two-peptide" rule for protein identification is an unsubstantiated heuristic that often results in loss of a large number of protein identifications (20%-25% of all expressed proteins).

Another unsubstantiated assumption often made by proteomics researchers is that maximizing the number of peptide identifications automatically results in maximizing the number of protein identifications. While peptide and protein identification goals are closely related, there are cases when one is more important than the other. Optimizing protein-level FDR is critical in applications such as biomarker discovery or proteome profiling, while optimizing peptide-level FDR is important for label-free quantitations [86], proteogenomics [53, 51], or peptidomics [15, 39]. We demonstrate that peptide and protein identifications are two different computational problems that should be approached differently: maximizing the number of peptide identifications does not necessarily result in maximizing the number of protein identifications (for example, in cases when most peptides hit a few proteins).

While FDR among *all* protein identifications in a dataset can be computed using the decoy database, computing the false positive rate (FPR) of *individual* protein identifications has been an open problem. The existing tools provide probabilistic scores, which may be correlated with the FPR, but do not provide its rigorous estimate taking into account the lengths of the proteins or the size of the spectral datasets. We extended the generating function framework [75, 74] to suggest a different (and simple) approach to the problem of estimating the protein-level FPR. So far, we failed to find any evidence that the previously proposed techniques for evaluating protein identifications (often based on elaborate machine learning models and multiple peptides) improve over a simple "single-peptide" rule combined with the generating function approach.

2.2 Methods

MS/MS Datasets. The MS/MS datasets used in this study were obtained from *Shewanella oneidensis* MR-1 (generated in Dick Smith's lab at PNNL) and human (generated in Vivian Hook's lab at UCSD) samples. These datasets are described in [53, 51, 50]. The datasets were generated on Thermo LCQ mass-spectrometer for *Shewanella* and Agilent XCT Ultra for human samples. The human dataset includes nearly 600 thousand spectra, while the *Shewanella* dataset has 14.5 million spectra. The *Shewanella* dataset was searched against the *Shewanella* protein database (size $\approx$ 1.5 MB, containing 4,928 sequences), and the human dataset was searched against the IPI database version 3.41 (size $\approx$ 40 MB, containing 72,155 sequences). Decoy databases were generated by randomly shuffling the sequence of each protein in the target database (preserving the background amino acid frequencies for each protein).

Peptide Identification. Database searches were carried out using InsPecT [126] and X!Tandem [26]. InsPecT searches were run with the default parameter settings (fragment ion tolerance of 0.5 Da and parent mass tolerance of 2.5 Da). X!Tandem was run using default settings, without any protease specificity (allowing peptides of length up to 40). MS-GeneratingFunction [75, 74] was run on InsPecT results to evaluate the statistical significance of individual peptide identifications (*spectral probabilities*). We treat the spectral probability as a score, and since MS-GF was used to rescore all InsPecT identifications (including even the very low scoring ones that are typically never reported), InsPecT$\oplus$MS-GF can be viewed as the third peptide identification tool, besides InsPecT and X!Tandem. MS-GeneratingFunction (MS-GF) was recently shown to improve upon InsPecT, X!Tandem and Sequest/PeptideProphet [75, 72].

Protein Identification. Protein identifications are inferred by applying score-thresholds to the peptides identified in a protein. In case of the “one-peptide” rule (i.e. allowing one-hit-wonders), any protein that has a peptide scoring above the chosen threshold is considered identified. This is equivalent to using a protein-level scoring scheme where the score of a protein is computed as the score of its highest scoring peptide. Similarly, in case of “two-peptide” rule, any protein that has two or more peptides scoring above the threshold is considered identified. Each point on the ROC curves is generated by changing this threshold, and computing the number of protein identifications by each rule, in the target and the decoy databases.

ProteinProphet was run using the Trans Proteomics Pipeline, v4.0 JETSTREAM rev 2, to process X!Tandem search results on human dataset. Both target and decoy protein sequences were included in the X!Tandem search results (without giving this information to ProteinProphet a priori). The final output thus included both target and decoy proteins with a probability-based score for each protein. The ROC curve was computed by varying the value of this score-threshold between 0 and 1. ProteinProphet performs better than the traditional “two-peptide” rule (as expected), but still worse than the one-peptide rule.

2.3 Results

2.3.1 Comparison of two-peptide and single-peptide rule

The intuition behind the “two-peptide” rule is that it may result in a more severe penalty to decoy hits as compared to the target hits. In other words, removing one-hit-wonders should improve the FDR of peptide and protein identifications. Figure 2.1 shows this trend for peptide identifications in the *Shewanella* dataset for InsPecT and MS-GF scoring functions. Indeed, if we discard all peptides representing one-hit-wonders, we observe that the tradeoff between the number of *peptides* identified in the target and the decoy database for different score thresholds shifts in favor of the target hits, compared to the situation when one-hit-wonders are retained.

Common sense suggests that the increased number of peptides, for a given FDR, should also increase the number of protein identifications. Therefore, it seems plausible that the “two-peptide” rule (discarding one-hit-wonders) should perform better than the “single-peptide” rule (retaining all proteins with one or more peptides). Note that in common practice, the “two-peptide” rule is always used in the context of protein identifications, i.e. after the peptides have been identified for a chosen FDR at peptide level.

Figure 2.2 demonstrates that this intuitive conclusion is not substantiated by the data as the simpler “single-peptide” rule has superior FDR as compared to the traditional “two-peptide” approach applied to protein identifications. For example, for the same number (21) of proteins identified in the decoy database, InsPecT [126] identifies 546 proteins in the target human database using the standard “two-peptide” approach and 742 proteins using the “single-peptide” approach, a 36% increase in the number of protein identifications. Similarly, the number of protein identifications with X!Tandem [26] increases from 350 to 414 (Figure 2.3), while the number of protein identifications with MS-GF [75] increases from 607 to 826. Similar trends are seen for the *Shewanella* datasets (Figure 2b).

Besides the “two-peptide” rule, more complex approaches are sometimes used to combine evidence from multiple peptides into protein identifications [67,97,41,94,143]. We compare our results with ProteinProphet [97], a popular tool that can be used to post-process MS/MS database search results of programs like Sequest and X!Tandem. Benchmarking ProteinProphet (combined with X!Tandem) against the “single-peptide” approach (Figure 2.3) shows that the “single-peptide” rule has better sensitivity specificity tradeoff than ProteinProphet, in both human and *Shewanella* datasets. This surprising result indicates that the simple “single-peptide” rule should not be discarded without benchmarking it against seemingly more reasonable (“two-peptide” rule) and complex (ProteinProphet) approaches.

We note that computing FDR using decoy databases (at the protein level) has some limitations because the number of protein identifications in the decoy database may deviate from the number of incorrect protein identifications in the target database, particularly in datasets with high proteome coverage. However, the relative comparison of the “single-peptide” and the “two-peptide” approaches should not be significantly affected by this phenomenon. The next section describes an approach for computing the protein-level error-rates without using a decoy database.

Inference of proteins from identified peptides is complicated by the presence of peptides that are present in multiple proteins. This problem is more pronounced in eukaryotes (since their proteomes have many repeated peptides) as compared to bacteria that have very few repeated peptides. The analysis of *Shewanella* in Gupta et al., 2007 [53] revealed that 98.5% of the identified peptides were unique, but the number can drop down to 40-50% in case of higher eukaryotes when using a sequence database like the human IPI database containing alternatively-spliced variants of proteins. A number of approaches have been developed to address this problem, most of which are parsimony-based [97, 96, 3, 81]. The computation of FDR in our method can depend on the exact method used for protein inference. For the sake of simplicity, we assign the ambiguous peptides (that map to multiple proteins) to one of the matching proteins randomly. To check that our results are robust to this step, we tried limiting our attention to only the unique peptides that are identified by InsPecT in the human dataset. We still find that the "single-peptide" approach works best for protein identifications (Figure 2.4). This shows that our conclusions are not significantly altered by peptides matching multiple proteins.

2.3.2 Estimating statistical significance of protein identifications using spectral dictionaries

We say that a protein matches a spectrum with score $Score$ if one of its peptides matches the spectrum (with the same $Score$). Consider the following two problems:

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

Protein-Spectra Matching Problem. Given a spectral dataset $Spectra$ and a score threshold $Threshold$ for a spectrum-peptide scoring function, find the probability that a random protein matches a spectrum in the set $Spectra$ with score equal to or larger than $Threshold$.

Recently, Kim et al., 2008 [75] suggested a generating function approach (MS-GF) for solving the Peptide-Spectrum Matching Problem and thereby evaluating the False Positive Rates (FPR) of Peptide-Spectrum Matches (PSM). In difference from

False Discovery Rates (FDR) empirically computed by the target-decoy approaches (that lack the ability to evaluate the statistical significance of *individual* PSMs), MS-GF rigorously evaluates individual PSMs using *spectral probabilities*. However, Kim et al., 2008 [75] defined the spectral probabilities for a *single* peptide and spectrum, while MS/MS searches compare *multiple peptides* (e.g., all peptides present in a protein database) against *multiple* spectra. Therefore, accurate estimation of protein-level FPRs requires a solution to the yet unsolved Protein-Spectra Matching Problem. In this section, we address this problem by extending the generating function framework from ***Peptide-Spectrum*** matches to ***Protein-$\mathcal{S}pectra$*** matches.

We start by reviewing the terminology related to spectral probability and spectral dictionaries [75, 74] (see Table 2.1). Given a scoring function for PSMs, $Dictionary(Spectrum, Threshold)$ is defined as the set of all possible peptides with $Score(Peptide, Spectrum) \geq Threshold$. The spectral dictionary framework transforms the (difficult) problem of evaluating the statistical significance of a PSM with score $Threshold$ into a (simple) problem of evaluating the statistical significance of matches between a set of strings ($Dictionary$) and a random peptide. Assuming probabilities of all amino acids in a random peptide equal to $\frac{1}{20}$, the probability that a $Dictionary$ contains a random peptide (more precisely, a prefix of a sufficiently long random string of amino acids) is given by

$$SpectralProbability(Dictionary) = \sum_{Peptide \in Dictionary} 20^{-|Peptide|}$$

where $|Peptide|$ stands for the length of $Peptide$. The dictionary corresponding to a PSM ($Dictionary(Peptide, Spectrum)$) is defined as $Dictionary(Spectrum, Score(Peptide, Spectrum))$. Similarly, the spectral probability of a PSM is defined as

$$SpectralProbability(Peptide, Spectrum) =$$

$$SpectralProbability(Dictionary(Peptide, Spectrum))$$

This allows one to convert an arbitrary (additive) scoring function $Score(Peptide, Spectrum)$ into a new scoring function represented by $SpectralProbability(Peptide, Spectrum)$. Kim et al, 2008 [75] demonstrated that this new scoring function (varying between 0 and 1) results in a better sensitivity-specificity trade-off than all other scoring functions they evaluated. Note that in this new scoring function, the lower scores (spectral probabilities) represent “better” PSMs. As such, given a $Spectrum$, one can define $Dictionary(Spectrum, SPThreshold)$, for a spectral probability threshold $0 \leq SPThreshold \leq 1$, as the set of all peptides with $SpectralProbability(Peptide, Spectrum) \leq SPThreshold$ when $SpectralProbability$ serves as a scoring function. We alert the reader that the term dictionary refers to both dictionaries defined by the original scoring function $Score$ and the new scoring function $SpectralProbability$. We further define $SpectralProbability(Spectrum, SPThreshold)$ as $SpectralProbability(Dictionary(Spectrum, SPThreshold))$. It is easy to see that the $SpectralProbability$ score (in contrast to the original $Score$) satisfies the important property:

$$SpectralProbability(Spectrum, SPThreshold) = SPThreshold$$

In the remainder of this work, we assume that $SpectralProbability$ is the scoring function being used.

The probability that a $Dictionary$ contains a peptide from a random protein of length n can be computed as $1 - \overline{P}(Dictionary, n)$, where $\overline{P}(Dictionary, n)$ is the probability that none of the peptides in the $Dictionary$ is contained in a random string of length n. Computing $\overline{P}(Dictionary, n)$ is a non-trivial problem that was solved by Guibas and Odlyzko, 1981 [49] using the generating function approach. While this approach allows one to compute $\overline{P}(Dictionary, n)$ precisely, the resulting expressions and recurrences are rather difficult to analyze due to *correlations* between different strings in the $Dictionary$ (see [49] for details). We therefore prefer to use an approximation (that ignores correlations), a reasonable assumption for a rather large alphabet of 20 amino acids (the effect of correlations is reduced with the increase in the alphabet size [49]). Under this assumption, $\overline{P}(Dictionary, n)$ is approximated as $(1 - SpectralProbability(Dictionary))^n$.

Therefore, the probability that a $Dictionary$ contains a peptide from a random protein $Database$ can be approximated as

$$1 - (1 - SpectralProbability(Dictionary))^{|Database|} \approx$$

$$SpectralProbability(Dictionary) \cdot |Database|$$

(under the condition that $SpectralProbability(Dictionary) \cdot |Database| << 1$) [75]. This condition is satisfied in practice since otherwise one ends up with spectral identifications that feature an unacceptably high FPR. Since we can represent a PSM by a $Dictionary(Peptide, Spectrum)$, the probability of a Protein-Spectrum Match (with a random protein) can be similarly computed as $SpectralProbability(Dictionary) \cdot |Protein|$. Therefore, if one accepts peptides with spectral probability score $SPThreshold$ and below ($0 \leq SPThreshold \leq 1$), then the probability of a match with a random $Protein$ (spectral probability of Protein-Spectrum Matches) is approximated as $SPThreshold \cdot |Protein|$.

Accounting for protein length

The score of a protein identification was previously computed as the score of its best scoring PSM, without taking into account the length of the protein. This is an over-simplification since longer proteins are more likely to have spurious matches than shorter ones. The spectral probability of Protein-Spectrum Matches ($SpectralProbability(Dictionary) \cdot |Protein|$) suggests a natural normalization for computing protein-level FPR (note that while this normalization is applicable to the $SpectralProbability$ score, it is not necessarily valid for other scoring functions). Figure 2.5 compares this Protein-Spectrum score (normalized by the protein length) with the original Peptide-Spectrum score(spectral probability before normalization) and show that the proposed normalization makes sense (for both human and *Shewanella* datasets). While this normalization does not result in a large change in the sensitivity-specificity tradeoff (since most proteins have similar lengths), the normalized Protein-Spectrum scoring function does show a modest increase in the number of protein identifications in the target database. For example, for 20 proteins identified in the decoy database (human dataset), the Peptide-Spectrum scoring identifies 838 proteins in target database while the Protein-Spectrum scoring (normalized by the protein length) identifies 865 proteins.

2.3.3 Accounting for the size of spectral dataset

We now extend the spectral dictionary approach from [74] to evaluating the statistical significance of matches between entire spectral dataset $\mathcal{S}pectra$ and a protein. Like protein length, the size of the spectral dataset also affects the statistical significance of a protein identification. While large spectral datasets are more likely to produce spurious identifications, some protein identification tools do not correct for the spectral dataset size. When computing the probability of a match below the score threshold $0 \leq SPThreshold \leq 1$ for a *single* spectrum (we remind the reader that smaller $SpectralProbability$ scores correspond to better Peptide-Spectrum matches), one could correct for the size of the $Protein$ as $SPThreshold \cdot |Protein|$. However, when estimating the number of hits for *multiple* spectra, a similar correction $SPThreshold \cdot |\mathcal{S}pectra|$ does not apply because of the correlations between the spectra within a typical spectral dataset (e.g., multiple spectra of the same peptide). To address this problem, we introduce the notion of a *spectral lexicon* below. We start by defining the dictionary of a spectral dataset as

$$Dictionary(\mathcal{S}pectra, Threshold) =$$

$$\bigcup_{Spectrum \in \mathcal{S}pectra} Dictionary(Spectrum, Threshold)$$

A string from $Dictionary$ is called *redundant* if its proper substring also belongs to $Dictionary$ (e.g. PEPTIDE is redundant if PEPTID, or PTIDE, or any of the shorter substrings belong to the dictionary). We define
$Lexicon(Dictionary(\mathcal{S}pectra, Threshold))$, represented more succinctly as
$Lexicon(\mathcal{S}pectra, Threshold)$, as the set obtained by removal of all redundant strings from $Dictionary(\mathcal{S}pectra, Threshold)$. While $Lexicon(\mathcal{S}pectra, Threshold)$ combines dictionaries of all spectra in the dataset, for all practical purposes it can be treated as a dictionary of a single (virtual) spectrum. Moreover, the probability that a $Lexicon$ contains a random peptide (i.e., that a spectrum from $\mathcal{S}pectra$ matches a random peptide with a score equal to or better than $Threshold$) is again given by
$SpectralProbability(Lexicon)$. Similarly, the probability that a $Lexicon$ contains a peptide from a random $Protein$ (i.e., the probability of a Protein-$\mathcal{S}pectra$ Match) can

be approximated as

$$1 - (1 - SpectralProbability(Lexicon))^{|Protein|} \approx$$

$$1 - e^{-SpectralProbability(Lexicon) \cdot |Protein|}$$

. Again, if $SpectralProbability(Lexicon) \cdot |Protein| << 1$, this probability can be approximated as $SpectralProbability(Lexicon) \cdot |Protein|$. We remark that while the expression $1-(1-SpectralProbability(Lexicon))^{|Protein|}$ is an approximation (the exact formula is given in [49]), it nevertheless leads to a reasonable estimate of Protein-$\mathcal{S}pectra$ FPR in practice (see next section). While this simplified formula involves multiple levels of approximations, it remains reasonable for small values of spectral probability (such as 10^{-11} or lower), for the typical sizes of spectral datasets (10^6) and protein lengths (10^3). For analyzing larger datasets, we recommend using more stringent (smaller) thresholds for spectral probability.

While $SpectralProbability(Lexicon)$ evaluates the statistical significance of Protein-$\mathcal{S}pectra$ matches (protein-level FPR), it remains unclear how to compute it. To address this problem, we define the notion of $Compression$ of a spectral dataset as a way to analyze dependencies between spectra. Loosely speaking, $Compression$ is the ratio of the spectral probability of the lexicon (of a spectral dataset) to the sum of spectral probabilities of individual dictionaries (of each spectrum). The ratio will often be less than 1 due to removal of redundant peptides from the lexicon and to the fact that many individual dictionaries share the same peptides. More precisely,

$$Compression(\mathcal{S}pectra, SPThreshold) =$$

$$\frac{SpectralProbability(Lexicon(\mathcal{S}pectra, SPThreshold))}{\sum_{Spectrum \in \mathcal{S}pectra} SpectralProbability(Dictionary(Spectrum, SPThreshold))}$$

Since we use $SpectralProbability$ as the scoring function,

$$SpectralProbability(Dictionary(Spectrum, SPThreshold)) = SPThreshold$$

and therefore,

$$Compression(\mathcal{S}pectra, SPThreshold) =$$

$$\frac{SpectralProbability(Lexicon(\mathcal{S}pectra, SPThreshold))}{|\mathcal{S}pectra| \cdot SPThreshold}$$

Two spectra are called ***independent*** if their dictionaries do not overlap and the union of their dictionaries does not contain redundant peptides. If all spectra in the spectral dataset were independent, then $Compression = 1$ and the quantity $|\mathcal{S}pectra| \cdot SPThreshold$ would provide a rigorous solution to the problem of the statistical significance of ***Peptide-$\mathcal{S}pectra$*** Matches. In reality, however, spectra of the same and related peptides (e.g., peptides that represent subpeptides of other peptides) that are typically present in the spectral dataset reduce $Compression$. While it can be explicitly computed for small spectral datasets, its efficient evaluation for large spectral datasets remains an open problem.

Spectral compression and spectral clustering

To estimate $Compression$ of a spectral dataset, we selected all high-accuracy (FT) *Shewanella* spectra with parent mass varying between 1090 and 1100 Da resulting in the dataset $\mathcal{S}pectra$ with 12,617 spectra. For different values of $SPThreshold$, $Dictionary(Spectrum, SPThreshold)$ were generated for each spectrum in $\mathcal{S}pectra$ and combined into $Lexicon(\mathcal{S}pectra, SPThreshold)$. Since all peptides in these dictionaries have similar parent mass (and thus do not have redundant peptides), $Lexicon$ can be generated by simply taking a union of spectral dictionaries of individual spectra. We find that $Compression$ values vary between between 0.65 and 0.89 depending on the spectral probability threshold (Table 2.2). This experiment indicates that while the expression $|\mathcal{S}pectra| \cdot SPThreshold$ over-estimates FPR, it is still within $\approx 65\% - 89\%$ of the correct estimate, a reasonable approximation.

Below we describe an alternative approach to estimating $Compression$ via *spectral clustering*. We assume that a spectral sample $\mathcal{S}pectra$ represents the set of peptides $\mathcal{P}eptides$. While the set $\mathcal{P}eptides$ is unknown, its size $|\mathcal{P}eptides|$ can be estimated by MS-Clustering tool [45] as the number of spectral clusters. Under the “ideal” scenario, the spectra of the same peptide (belonging to one cluster) have identical spectral dictionaries, spectra of different peptides do not overlap, and all spectra are independent.

Under these assumptions,

$$SpectralProbability(Lexicon(\mathcal{S}pectra, SPThreshold)) = |\mathcal{P}eptides| \cdot SPThreshold$$

Therefore,

$$Compression = \frac{|\mathcal{P}eptides| \cdot SPThreshold}{|\mathcal{S}pectra| \cdot SPThreshold} = \frac{|\mathcal{P}eptides|}{|\mathcal{S}pectra|}$$

In reality, dictionaries of spectra of the same peptides are not identical, dictionaries of spectra of different peptides may overlap, and spectra may be dependent. As a result, the parameter $\frac{|\mathcal{P}eptides|}{|\mathcal{S}pectra|}$ typically under-estimates $Compression$. For example, for the dataset $\mathcal{S}pectra$ consisting of spectra with parent mass between 1090 and 1100 Da, MS-Clustering [45] found 3,584 clusters from 8,689 spectra, as shown in Table 2.3 (3928 out of 12,617 spectra were discarded as low-quality spectra). The estimated $Compression = 3584/8689 \approx 0.41$ is lower than the values observed in Table 2.2. Table 2.4 shows that large clusters have $Compression$ much larger than $1/ClusterSize$ and illustrates that clustering under-estimates $Compression$.

While computing the FPR of Protein-$\mathcal{S}pectra$ matches enables one to evaluate the statistical significance of *individual* protein identifications, it also allows one to estimate the expected number of false identifications among all identifications, without requiring a decoy database. The following computational experiment shows that it is feasible to estimate the number of protein identifications in a decoy database without actually using a decoy database. The human spectra were searched with InsPecT⊕MS-GF against a database of 10,000 randomly generated proteins of length 500 each, with equal probability of each amino acid at each position. 286,717 spectra ($\mathcal{S}pectra$) got some InsPecT/MS-GF score against this database and were included in the analysis. Table 2.5 shows the *observed* number of proteins that have any peptide exceeding the score threshold ($SPThreshold$). The *expected* number of protein identifications, for different values of $SPThreshold$, is given by

$$10000 \cdot (1 - e^{-SpectralProbability(Lexicon) \cdot |Protein|})$$

where $SpectralProbability(Lexicon)$ is estimated as $Compression \cdot |\mathcal{S}pectra| \cdot SPThreshold$. This requires an estimate for the value of $Compression$ for the spectral dataset that we describe below.

Suppose that the spectral dataset $\mathcal{S}pectra$ is partitioned into non-overlapping $Clusters$ (i.e., $\mathcal{S}pectra = \bigcup_{Cluster \in Clusters} Cluster$), where each $Cluster$ represents the set of $|Cluster|$ spectra originating from the same peptide. If we assume that $Compression$ within each $Cluster$ is $1/|Cluster|$, then

$$SpectralProbability(Lexicon(\mathcal{S}pectra, SPThreshold)) = |\mathcal{C}lusters| \cdot SPThreshold$$

However, Table 2.4 reveals the limitations of this estimate. A more accurate estimate is given by:

$$SpectralProbability(Lexicon(\mathcal{S}pectra, SPThreshold)) \approx$$
$$\sum_{Cluster \in Clusters} SpectralProbability(Lexicon(Cluster, SPThreshold)) \approx$$
$$\sum_{Cluster \in Clusters} |Cluster| \cdot Compression(Cluster, SPThreshold) \cdot SPThreshold$$

Therefore,

$$Compression(\mathcal{S}pectra, SPThreshold) \approx$$
$$\frac{\sum_{Cluster \in Clusters} |Cluster| \cdot Compression(Cluster, SPThreshold)}{|\mathcal{S}pectra|}$$

While the parameter $Compression(Cluster, SPThreshold)$ can be explicitly computed for each cluster, Table 2.3 shows that this parameter depends largely on the cluster size (and $SPThreshold$). Therefore, we can approximate $Compression(Cluster, SPThreshold)$ by using the average value over the previously analyzed clusters of the same size. Using the cluster partitioning of the human dataset determined by MS-Clustering, the expected number of proteins was estimated using the above formula by plugging in average $Compression(Cluster, SPThreshold)$ values from Table 2.3. Table 2.5 compares this expected number of protein identifications with the number of proteins actually identified in the decoy database search and demonstrates that this approach allows one to get a reasonably close estimate of the number of false protein identifications without searching a decoy database.

2.3.4 Using FPR of protein identifications

This study recommends discarding the commonly used “two-peptide” rule and instead supports reporting protein identifications (including one-hit-wonders) according to their rigorous statistical significance (FPR). To illustrate this, Supplementary Table 1 provides a ranked list of protein identifications in the human dataset, sorted by increasing FPRs. Denote the FPR of the $i'th$ ranked protein as $FPR(i)$. Below we estimate the number of protein identifications in a decoy database (and thus FDR of protein identifications) if only proteins with FPRs equal to or exceeding $FPR(i)$ are reported.

If N is the total number of proteins in the sequence database ($N = 72155$ for the human IPI database used here), one can estimate the expected number of false positives among top i protein identifications as $False(i) = N \cdot FPR(i)$, and therefore, the estimated FDR at that stringency level is $FDR(i) = \frac{\#proteins\ in\ decoy\ database}{\#proteins\ in\ target\ database} = \frac{N \cdot FPR(i)}{i}$. As i is increased from 1 to ≈ 600 in this dataset, $False(i)$ increases from 0 to 0.2 only, indicating that these top 600 protein identifications are almost error-free. It is worth noticing that 160 of these extremely reliable protein identifications are one-hit-wonders, a large fraction that may be missed by the traditional approaches favoring multiple peptides. When i is increased from 600 to 800, $False(i)$ increases to 21, indicating that a tenth ($\approx 21/200$) of the protein identifications in this range may be incorrect. Increasing i beyond 800 rapidly increases $False(i)$, eventually at a rate higher than the rate of increase of i (see the last few rows in the table), thereby essentially reducing the number of true identifications. Therefore one can choose to select the top ≈ 800 proteins in this list (at an FDR of $21/800 \approx\ 2.5\%$) as a reasonable set of protein identifications. This analysis shows how the knowledge of protein-level FPRs allows one to make informed judgement calls in selecting reliable protein identifications from many hits obtained in database searches.

2.4 Discussion

We have demonstrated that the commonly used “two-peptide” rule jeopardizes the sensitivity-specificity trade-off in protein identifications. This counter-intuitive observation points out that we are more likely to get a protein with two mediocre peptide hits in the decoy database (by chance) than a single high-scoring peptide hit. Some software tools (e.g. ProteinProphet([97]), Panoramics([41]) identify proteins if the combined score from all peptides exceeds the threshold, and thus allow scoring some single-hit proteins higher than proteins with multiple hits. However, these approaches are dependent on specific scoring models and require further theoretical justification to be suitable for use with all protein identification tools. In particular, our results indicate that the simpler “single-peptide” rule results in better sensitivity/specificity trade-off than ProteinProphet.

One might expect that since proteins are inferred from peptides, optimizing peptide identifications also optimizes protein identifications. We have provided evidence that this intuition does not hold ground. In reality, when one attempts to maximize the number of peptides in the target database, the additional peptides often come from the already covered proteins (i.e., from proteins with more than 1 match) and thus do not increase the overall number of protein identifications. However, the corresponding increase in the number of peptides in the decoy database significantly raises the number of decoy protein identifications (and thus increases FDR). Therefore, we emphasize that optimizing FDRs for peptides and proteins must be considered as different problems that are best addressed by different approaches.

This study does not recommend accepting *all* proteins with single peptide hits but instead argues that the “single-peptide” approach must be used in conjunction with control of the FDR. While it may be surprising that the “single-peptide” approach generates a larger set of identified proteins than the seemingly more reliable “two-peptide” approach (without sacrificing FDR), our results indicate that it is the case. We demonstrated that for any set of proteins identified by the “two-peptide” approach (with peptide score threshold x), there is a larger set of protein identified by the “single-peptide” approach with the same FDR (with a more stringent peptide score threshold $x + \epsilon$). Therefore, one has to choose the peptide-level score thresholds carefully to ensure that the “single-peptide” approach and “two-peptide” approach are being used for the same level of FDR.

We also discussed how to estimate the FPR of individual Protein-Spectra matches using the generating function framework. Spectral clustering appears as a promising albeit an indirect approach for approximating spectral compression and should be explored further.

While the proteomics community often takes great care in evaluating peptide-

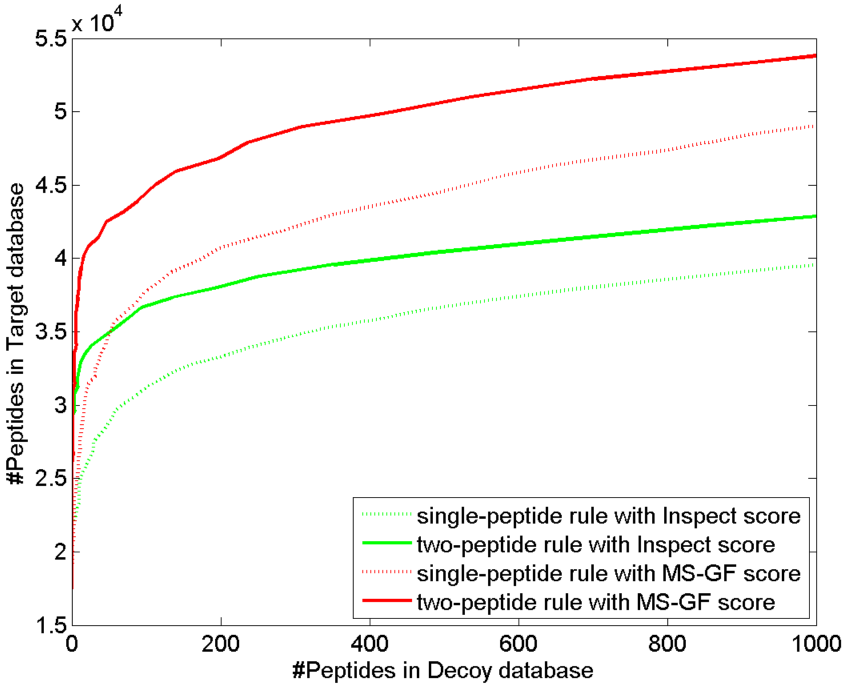
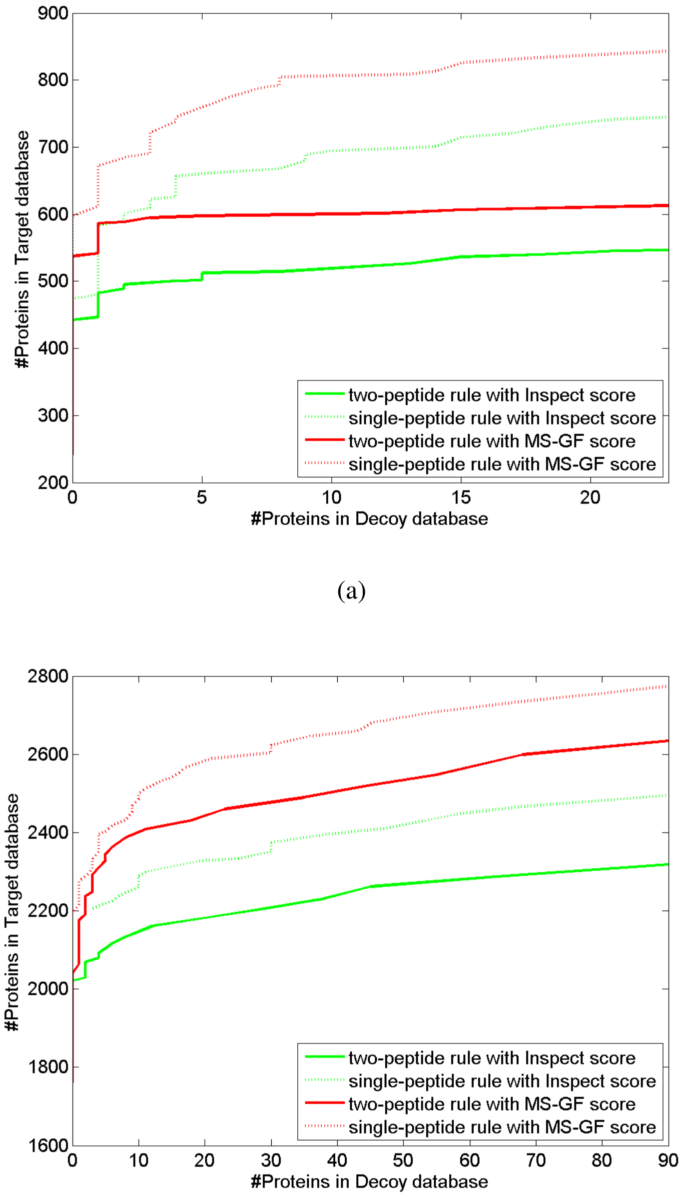


Figure 2.1: Identification of peptides in the *Shewanella* dataset using different approaches and scoring functions. Each point in the curves is generated by varying the scoring threshold and computing the number of hits in the target and the decoy database exceeding the threshold.

level error rates, the protein-level FDRs and FPRs are rarely computed. We suggest that publications reporting protein identifications should also report protein-level FPRs and/or FDRs. Lack of this checkpoint raises concerns about the validity of studies, such as biomarker discovery, where the number of identified proteins as well as the reliability of each individual identification is of critical importance.

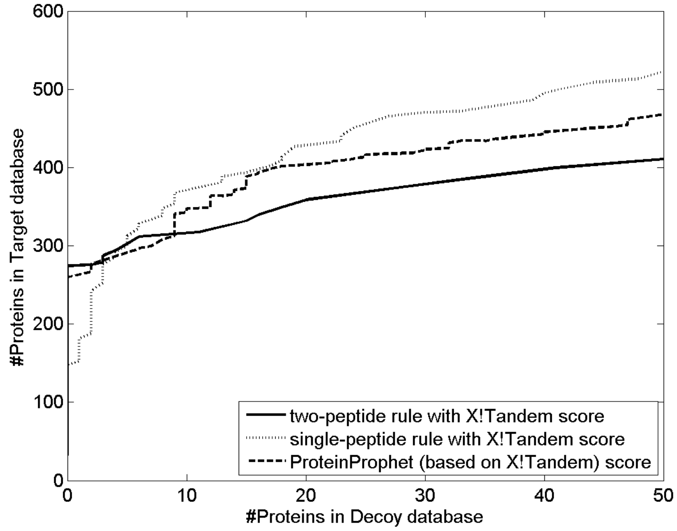
Chapter 2 is, in part, a reprint of the paper "False discovery rates of protein identifications: a strike against the two-peptide rule. N. Gupta and P.A. Pevzner (2009). Journal of Proteome Research. 8(9):4173-81". The dissertation author was the primary investigator and author of this paper.



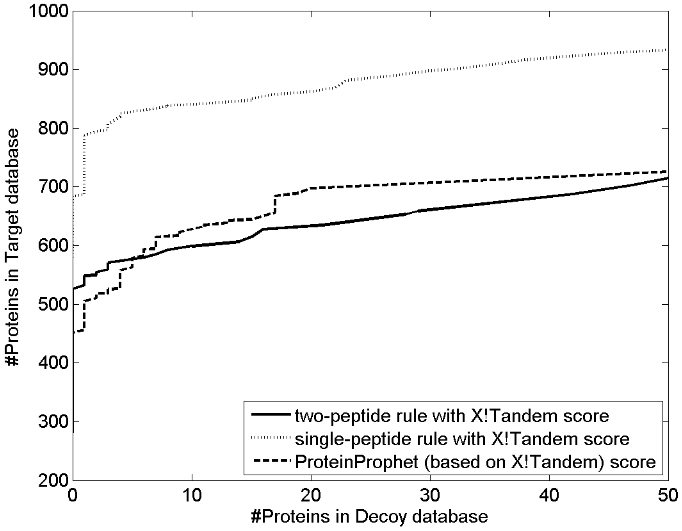
(a)

(b)

Figure 2.2: **(a)** Identification of proteins in the human dataset using different approaches and scoring functions. **(b)** Similar plot as in (a) for *Shewanella* dataset.



(a)



(b)

Figure 2.3: **(a)** Protein identification in the human dataset using X!Tandem search results with different scoring approaches at the protein level. **(b)** Similar plot as in (a) for an arbitrarily selected subset of *Shewanella* dataset containing 1.25 million spectra.

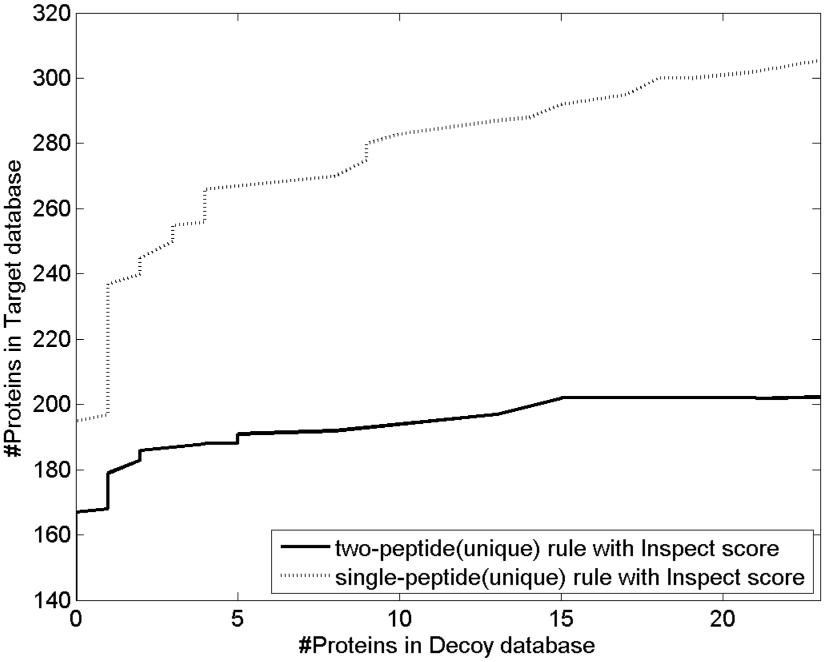
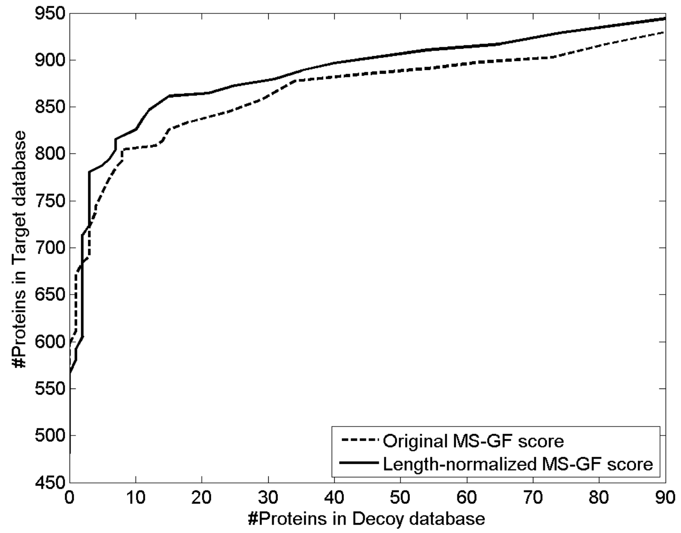
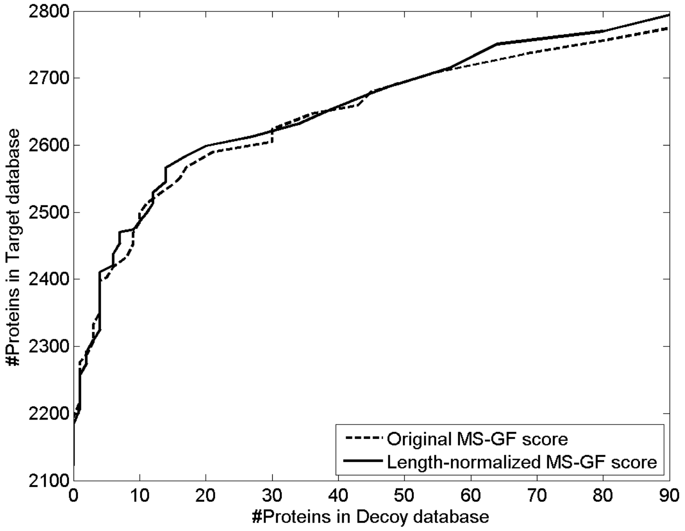


Figure 2.4: Identification of proteins, using the unique peptides only (peptides that are not shared between multiple proteins), in the human dataset using InsPecT search results with different approaches.



(a)



(b)

Figure 2.5: **(a)** Identification of proteins in the human dataset using MS-GF scores, without and with length correction. **(b)** Similar plot as in (a) for *Shewanella* dataset.

Table 2.1: An intuitive description of some frequently-used terms (please refer to the text for details).

Term	Description
$Dictionary(Spectrum, Threshold)$	A set of peptides that match the $Spectrum$ with scores greater or equal to the $Threshold$
$SpectralProbability(Dictionary)$	The probability that a random peptide matches a sequence contained in the $Dictionary$
$Lexicon(\mathcal{S}pectra, Threshold)$	Union of the dictionaries of all spectra in the set $\mathcal{S}pectra$ (for the given $Threshold$), after removing redundant peptides
$Compression(\mathcal{S}pectra, Threshold)$	Ratio of the spectral probability of the $Lexicon$ to the sum of the spectral probabilities of each spectral dictionary $Dictionary(Spectrum, Threshold)$ for all spectra from the set $\mathcal{S}pectra$

Table 2.2: $Compression$ for different values of spectral probability threshold ($SPThreshold$) on a small spectral dataset ($\mathcal{S}pectra$) consisting of 12,617 spectra, with parent mass between 1090 and 1100 Da. The second column represents the sum of the sizes of the dictionaries of individual spectra and the third column represents the size of the combined $Lexicon$.

$SPThreshold$	$\sum_{S \in \mathcal{S}pectra} \lvert Dictionary(S, SPThreshold) \rvert$	$\lvert Lexicon(\mathcal{S}pectra, SPThreshold) \rvert$	$Compression$
1e-12	444027	314347	0.752
1e-11	2709887	1708317	0.737
1e-10	16919910	9782247	0.689

Table 2.3: Summary of spectral clusters produced by MS-Clustering on the set of 12,617 spectra with parent mass between 1090 and 1100 Da. 8,689 spectra were clustered into 3,584 clusters, while the remaining 3,928 spectra were discarded by MS-Clustering due to low spectral quality. The last three columns indicate the average values of $Compression$ for various cluster-sizes, for three different values of spectral probability threshold ($SPThreshold$).

			Average $Compression$		
Cluster-size	# Clusters	# Spectra	1e-10	1e-11	1e-12
1	2875	2875	1	1	1
2	234	468	0.96	0.97	0.98
3	115	345	0.88	0.90	0.92
4	71	284	0.87	0.88	0.90
5	47	235	0.83	0.93	0.95
6	36	216	0.80	0.86	0.88
7	24	168	0.76	0.81	0.85
8	20	160	0.75	0.84	0.95
9	17	153	0.66	0.73	0.87
10	17	170	0.68	0.80	0.84
10-15	40	506	0.71	0.81	0.83
15-20	27	475	0.60	0.65	0.75
21-30	23	584	0.47	0.55	0.65
31-40	11	380	0.48	0.65	0.77
41-50	8	370	0.28	0.41	0.50
>50	19	1300	0.23	0.47	0.45
Total	3584	8689			

Table 2.4: Compression computed for the ten largest clusters in Table 2.3, using dictionaries generated with the spectral probability threshold of 1e-10.

Rank (by size)	Size	$Compression$
1	125	0.246
2	82	0.211
3	77	0.128
4	75	0.126
5	74	0.080
6	72	0.495
7	72	0.066
8	71	0.235
9	70	0.237
10	66	0.361

Table 2.5: Comparison between the observed and the expected number of protein identifications for different values of spectral probability threshold ($SPThreshold$), when searching the human spectral dataset against a randomly generated decoy database consisting of 10,000 proteins of size 500 each. The values of $Compression$ were estimated by using the average values for different cluster-sizes from Table 2.3.

$SPThreshold$	$Compression$	Expected # IDs	Observed # IDs
1e-10	0.88	125.36	80
1e-11	0.91	13.04	14
1e-12	0.92	1.32	1

Chapter 3

Proteogenomics

3.1 Introduction

Gene annotation continues to be a challenging task requiring both automated analysis and manual curation. Even in a seemingly simple case of bacterial gene prediction, many challenges remain and a large number of genes are annotated incorrectly or even missed. This is mainly due to difficulties in prediction of short genes, genes with unusual codon usage, as well as accurate prediction of Start codons. This challenge is greatly magnified in recent meta-genomic projects, which seek to sample DNA from the environment.

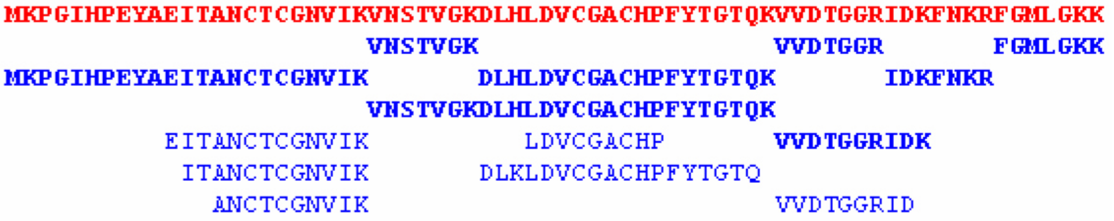
The goal of proteogenomic annotations is to use mass spectrometry data for annotating genome and use the resulting genomic annotations to improve MS-based protein identification and the proteome annotation. We argued that complementing sequencing projects by MS/MS projects would significantly improve both genome and proteome annotations. However, to make this happen, automated software pipeline for proteogenomic annotation had be developed and integrated with the existing gene prediction tools. We emphasize that proteogenomic annotations go well beyond gene finding and include signal peptide predictions, RNA recoding predictions, operon predictions, etc.

The idea of querying MS/MS dataset against a genome to identify protein coding genes has been used earlier in different settings [148, 78, 103, 65, 64, 140, 69, 42]. However, the problem is far from being simple and there are still no program that can automatically compare a large MS/MS dataset against a genome and to come up with a list of new and corrected gene annotations. While such program would be of great interest to genomics community, designing such program even for bacterial (let alone eukaryotic) genome remains a challenge. Even more challenging is the problem of annotating important regulatory mechanisms like programmed frameshifts that currently remain beyond the reach of gene prediction algorithms. While our preliminary proteogenomic analysis of *Shewanella oneidensis* showed that mass-spectrometry can address these important problems, many computational challenges remain in using MS/MS for gene annotation and analyzing post-translational processing (e.g., proteolysis).

A major limitation in the development of this field was that until recently, complex genome-wide MS/MS searches(like search for mutations/polymorphisms or unrestricted search for modifications) were not feasible since the search for mutated/modified peptides was prohibitively time-consuming. The search becomes particularly time-consuming in the case of non-tryptic peptides that enable identification of proteolytic events. Moreover, our analysis revealed at least 24 modification types present in the *Shewanella oneidensis* sample, a significantly larger number than the existing restricted PTM search tools can handle under realistic parameters. Inspect/MS-Alignment removed this bottleneck and allowed us to generate accurate genome-wide proteogenomic annotations.

3.2 Results

Dick Smith's lab (PNNL) generated a dataset of 14.5 million tandem mass spectra for *S. oneidensis* for 17 cell culture conditions, the largest MS/MS dataset ever reported for a bacterium. *S. oneidensis* is an aero-tolerant anaerobe able to reduce heavy metal ions and remove them from solution, making it a model organism for bioremediation studies [95]. It has been extensively studied by the Shewanella Federation (http://shewanella.org/) and the predicted genes were manually validated in a number of studies. As a result, it is considered among the most well annotated bacterial genomes second only to *E. Coli*. Moreover, *Shewanella* is currently the only bacterial genome for which extensive comparative genomics information exists (multiple strands of *She-*



```
MKPGIHPEYAEITANCTCGNVIKVNSTVGKDLHLDVCGACHPFYTGTQKVVDTGGRIDKFNKRFGMLGKK
                       VNSTVGK                   VVDTGGR       FGMLGKK
MKPGIHPEYAEITANCTCGNVIK       DLHLDVCGACHPFYTGTQK       IDKFNKR
                       VNSTVGKDLHLDVCGACHPFYTGTQK
          EITANCTCGNVIK          LDVCGACHP       VVDTGGRIDK
           ITANCTCGNVIK       DLKLDVCGACHPFYTGTQ
             ANCTCGNVIK                          VVDTGGRID
```

Figure 3.1: The ribosomal protein L31 (SO_4120) is entirely covered by identified peptides. The protein sequence is shown at the top in red, and the identified peptides are shown below in blue. Tryptic peptides are shown in bold.

wanella have been sequenced) thus providing a possibility to further improve gene annotations. It therefore may come as a surprise that even in this well-annotated genome we found many new and mis-annotated genes. The difficulties in gene prediction are illustrated by the fact that just a year after *S. oneidensis* was first annotated, it was reannotated with a large number of changes [29]. Some recent studies have attempted to further improve the annotations via microarray and mass-spectrometry data [34, 35, 76, 113].

3.2.1 Using MS/MS data to find expressed proteins

Some previous proteogenomic studies [65, 64] did not attempt to measure the rate of false peptide identification, thus raising concerns about the reliability of new gene annotations. We searched 14.5 million MS/MS spectra against a six-frame translation (over 10 million amino acids) of the *S. oneidensis* MR-1 genome using InsPecT. A decoy database of shuffled sequences was created and searched as a negative control. In total 1.4M spectra were annotated while searching 14.5M spectra against the forward database.

We analyzed the locations of the identified peptides relative to the positions of the TIGR and GeneMark genes. Of the 331 peptides not covered by a TIGR gene, 126 were covered by GeneMark predictions. This demonstrates the utility of MS/MS data in resolving discrepancies between various gene prediction efforts.

We consider protein expression to be confirmed if at least two peptides were identified from that protein. Using this approach we verify expression of about half of the predicted proteome. We observed an excellent correlation of expressed proteins with both *S. oneidensis* predicted operons and various components of the cell machinery. For example, most proteins from a (i) ribosome, (ii) tRNA aminoacylation machinery, (iii) purine biosynthesis and (iv) central carbon metabolism are highly covered. A shift to lower coverage is illustrated by the biosynthesis of NAD cofactor and Lipid A, pathways that are expected to be relatively less active. A subsystem involved with the utilization of chitin and N-acetyglucosamine provides a remarkable example of correlation between the observed coverage and expected phenotype. Most of the genes recently implicated in the respective regulon in *S. oneidensis* [147] display no coverage, consistent with the notion that this nutrient, highly abundant in the natural marine habitat, was not a part of growth media used in this study. Notably, a predicted transcriptional repressor of this regulon was expressed.

3.2.2 Using MS/MS datasets to improve gene annotations

Multiple matches to a genomic region outside the boundary of genes can be used to detect new genes missed during genome annotation or to suggest that gene boundaries should be expanded.

To detect such cases, we examined the identified peptides falling outside the TIGR genes. Such peptides are combined into putative coding segments if they are located within a short distance and have compatible reading frames. These coding segments point to new genes and extensions of the TIGR genes. By analyzing the location of these segments, we identified eight new genes and extended the 5' boundaries for 30 genes. In difference from other proteogenomic studies (where closely related genomes were not available), we were able to evaluate the error rates of our re-annotations via comparative genomics approaches. It turned out that all newly proposed MS/MS-based annotations are confirmed by comparative genomics analysis.

Based on comparative sequence alignment, extensions of the 5' ends of all 30 genes were consistent with predicted N-termini of proteins identified in other bacteria. Figure 3.2 illustrates an example of how peptides were used to identify an earlier gene start position for SO_1175. Five frequently observed peptides (Figure 3.2(a)) align upstream to and within the same reading frame as the original SO_1175 gene (Figure

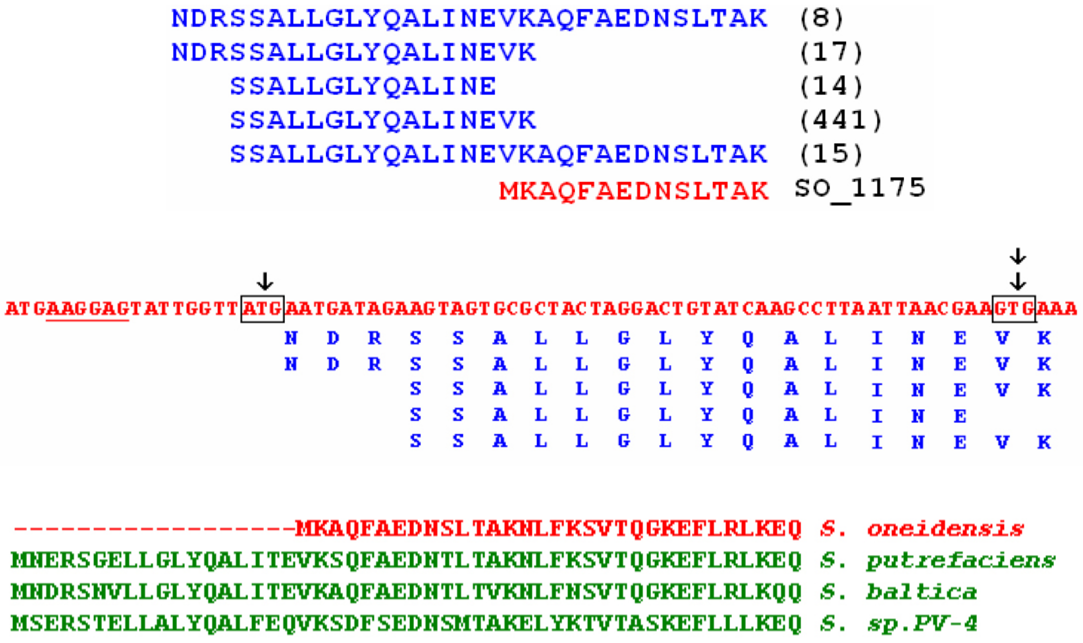

NDRSSALLGLYQALINEVKAQFAEDNSLTAK (8)
NDRSSALLGLYQALINEVK (17)
SSALLGLYQALINE (14)
SSALLGLYQALINEVK (441)
SSALLGLYQALINEVKAQFAEDNSLTAK (15)
MKAQFAEDNSLTAK SO_1175
ATGAAGGAGTATTGGTTATGAATGATAGAAGTAGTGCGCTACTAGGACTGTATCAAGCCTTAATTAACGAAGTGAAA
N D R S S A L L G L Y Q A L I N E V K
N D R S S A L L G L Y Q A L I N E V K
S S A L L G L Y Q A L I N E V K
S S A L L G L Y Q A L I N E
S S A L L G L Y Q A L I N E V K
------------------MKAQFAEDNSLTAKNLFKSVTQGKEFLRLKEQ S. oneidensis
MNERSGELLGLYQALITEVKSQFAEDNTLTAKNLFKSVTQGKEFLRLKEQ S. putrefaciens
MNDRSNVLLGLYQALITEVKAQFAEDNTLTVKNLFNSVTQGKEFLRLKQQ S. baltica
MSERSTELLALYQALFEQVKSDFSEDNSMTAKELYKTVTASKEFLLLKEQ S. sp.PV-4

Figure 3.2: (a) Alignment of identified peptides (colored blue) and the hypothetical TIGR protein SO_1175 (colored red). Numbers on the right show spectral counts The start codon is normally read as valine. These peptide identifications demonstrate that translation begins upstream of the annotated start site. (b) Alignment of identified peptides (blue) with the nucleotide sequence of N-terminal region of SO_1175, relative to proposed new start codon (single arrow) and to the original proposed start codon (double arrow). A Shine Dalgarno-like site (underlined) is found upstream only to the proposed new start site. (c) Multiple sequence alignment of SO_1175 of *S. oneidensis* (red) with orthologs in other *Shewanella* strains.

3.2(b)). The proposed new gene start is further validated by the detection of four peptides that span the original suggested start codon. Had this GTG codon occurred at the translational start position it would have been translated as a methionine rather than a valine [123]. As added proof, the alignment of SO_1175 with similar proteins deduced from other *Shewanella* genome sequences (Figure 3.2(c)) is consistent with the proposed new 5' extension of the gene.

Figure 3.3 illustrates how peptides upstream of SO_2300, which encodes translation initiation factor IF-3, led to the prediction of a non-traditional start codon (there is currently no gene prediction software capable of predicting non-traditional Start codons). Three peptides map upstream to and within the same reading frame as SO_2300 (Figure 3.3(a)). One of these peptides spans the original predicted GTG start codon, validating that it is translated as a valine rather than a methionine. However, none of the more common ATG, GTG, or TTG start codons was found upstream to the region spanned by these three peptides. Previous studies revealed that IF-3 gene from a variety of other bacteria is initiated at a rare ATT start codon thereby serving as a basis for autoregulation [114, 106, 60, 85]. While no ATT codon was found, we speculate that SO_2300 starts at an ATA codon (Figure 3.3(b)). ATA is known to function as a rare translation initiator [120, 123] and in this case is adjacent to a strong ribosome binding signal. One final line of evidence is shown in Figure 3.3(c) where results of Tblastn analysis reveal conservation in sequences spanning the entire proposed N-terminal extension of 11 *Shewanella* species as well as a large portion of this region in more distantly related bacteria.

3.2.3 Using MS/MS datasets to analyze mutations and modifications

MS-Alignment revealed over 4,000 mutations/modification sites with false discovery rate of 5% (many of these modifications can result from chemical damage in vitro [62]). Blind PTM search tools do not distinguish between chemical modifications and mutations with the the same DeltaMass. Some modifications correspond to amino acid substitutions, either due to polymorphisms, or due to errors in the genomic sequence. In collaboration with Margie Romine (PNNL) we validate such cases by considering the sequences of other *Shewanella* strains. For example, a modified peptide K.QQIG+14ENPIIVYMK.G from glutaredoxin domain protein can be explained by a glycine-to-alanine amino acid substitution. Indeed, analysis of the raw DNA sequence traces revealed a mistake in the genome sequence at the corresponding locus resulting in a GGT (Gly) rather than the correct GCT (Ala).

Application of mass spectrometry for whole-genome mutations or modification studies is a relatively unexplored territory. Even though we use the state-of-the-art methods for detection of chemical modifications, it remains a challenge to distinguish between in-vivo and in-vitro modifications, mutations, SAAPs, and sequencing errors. Some modifications types were observed on many different sites. Figure 3.4 presents 24 common modification types each observed on 5 or more distinct sites. Since the false positive rate is low it is extremely unlikely that any of these modification types represent a computational artifact. Moreover, all but two are known modification types, further reinforcing the conclusion that they are not artifact. We remark that the number of such modification types is rather large, significantly larger than the usual limit imposed by the restricted PTM search tools.

The current understanding of PTMs in bacteria is very limited even for well-studied organisms like *E. coli* or *Shewanella*. We anticipate that many biologically important PTMs in *Shewanella* and *E. coli* will be located on aligned positions in orthologous proteins. Indeed, several modifications we found appear on “orthologous” positions to those previously reported in *E. coli* (we remark that there are very few previously known modifications in *E. coli*). For instance, [77] reported the occurrence of β-methylthio-aspartic acid at D88 of ribosomal protein S12p in *E. coli* and suggested that it is important for stabilizing the ribosome structure. We observed the same modification at D89 of the *S. oneidensis* protein, the homologous position to D88 of the *E. coli* protein, as shown in Figure 3.6.

3.3 Identification of Signal peptides and N-terminal methionine cleavages

Signal peptide targets a protein for secretion, or for transportation to a desired cellular location. Signal peptides are cleaved and quickly degraded to produce the mature protein sequence. While knowledge of signal peptides is important for understanding protein function, they are difficult to confirm experimentally, and computational predictions are used to fill the gap. There have been some concerns [4] regarding the quality of popular signal peptide prediction algorithms (like SignalP [13] and PrediSi [57]) since these methods consider a generalized signal motif for all proteins and may not identify interesting cases that are limited to a few proteins. Also, these tools make predictions based on a rather small sample of experimentally confirmed signal peptides, since experimental data about signal peptides is limited. For example, SignalP makes predictions based on a dataset of only 334 experimentally confirmed signal peptides in all Gram-negative bacteria. The number of experimentally confirmed signal peptides in Gram-positive bacteria is twice smaller [13]. It is clear that MS/MS evidence can greatly increase the number of experimentally confirmed signal peptides and improve confidence of signal peptide predictions. However, there is still no computational tools for MS/MS-based prediction of signal peptides.

The N-terminal Methionine Cleavage(NME) is the process of cleaving Nterminal methionine residue by Methionyl Amino Peptidase (MAP) or Amino Peptidase P(AmpP) from a number of cytosolic proteins. NME is specific and one of the most common post-translational modification (it is estimated that $\approx 50\%$ of *E.Coli* proteins undergo PTM) with important implications for protein half-life [132]. While the knowledge of NME is crucial for many applications in food safety, infectious diagnostics, and counter-terrorism (it improves the quality of MS-based microorganism detection by an order of magnitude [30]), the computational algorithms for NME prediction remain somewhat simplistic.

NME is often crucial for function and stability of recombinant proteins [82]. Methionine, which is important during translation, may not be required (or actually be detrimental) for the function of the protein. The role of NME remains poorly understood but the process is recognized to be the major source of N-terminal amino acid diversity. The recognition rules for NME remain elusive resulting in a number of conflicting studies [30, 83, 142, 46]. Frottin et al., 2006 [46] recently estimated that the existing ambiguities in NME recognition rules make reliable proteome annotation difficult for about 30% of bacterial proteins. This renders the production of recombinant proteins of theraupitic interest risky, given the high antigenicity of the N-terminus if incorrectly processed, the problem originally encountered in the production of human growth harmone [12, 101].

In one of earlier NME studies, Hirel et al, 1989 [58] measured the efficiency of cleavage between initial methionine and various second residues in vitro and showed that methionine cleavage is more efficient if the second residue has a smaller side chain. The key limitation of this and follow-up studies of NME is the limited size of the experimental dataset. As a result, the software for predicting NME sites (TermiNator) was developed only recently [46]. We argue that analysis of large MS/MS dataset is able to generate orders of magnitude larger training datasets for studying NME and to resolve the problem of reliable NME annotation.

The above examples represent only two most well-studied cases of a myriad of proteolytic events characteristic for any organism. While regulatory proteolysis is crucial for many cellular processes, it remains poorly understood and there is still no high-throughput techniques for genome-wide detection of proteolysis. We argue that tandem mass spectrometry has a potential to become such technique as soon as computational tools for analyzing proteolytic events via MS/MS are developed.

Large MS/MS datasets offer an unprecedented opportunity to study in-vivo cleavage specificity by looking at non-tryptic peptides that may be manifestations of proteolytic events. If a protein sample is digested with trypsin, we expect the majority of the peptide endpoints to correspond to tryptic cleavage sites. Given trypsin's high specificity [100], it is natural to consider that non-tryptic endpoints may reveal proteolytic events [90]. Non-tryptic endpoints suggest the possibility of a proteolytic event, either in vivo or in vitro. In the *S. oneidensis* MR-1 project (collaboration with Dick Smith (PNNL)) we analyzed 14.5 million MS/MS spectra and arrived to peptide identifications that include 21,297 tryptic peptides (75%), 6,670 peptides with one non-tryptic endpoint

(24%), and 409 peptides with two non-tryptic endpoints (1%). However, caution is needed while analyzing peptides with non-tryptic endpoints, since they can also reflect post-digestion trimming of tryptic peptides or simply errors in peptide identifications. These peptides may create an appearance of proteolytic events that never happened thus calling for development of additional algorithmic and statistical analysis to distinguish between the true and false proteolytic events. Thus special care is needed to ensure that the rate of false positive identifications for non-tryptic peptides is low. Also, the whole-proteome search becomes rather time-consuming in the case of non-tryptic peptides that enable identification of proteolytic events. We note that 96% of non-tryptic peptides *S oneidensis* fall within expressed *S oneidensis* proteins, similar to tryptic peptides (97%). Since expressed proteins make up only 7% of the six-frame translation *S oneidensis*, incorrect identifications (which are randomly distributed) are unlikely to fall within confirmed proteins. Thus, we argue that our false discovery rate for non-tryptic peptides is not significantly larger than that for tryptic peptides.

3.3.1 Proteolytic events and non-covered peptides

Figure 3.7 shows the N-terminal portion of a well-covered *S.oneidensis* protein whose first 26 aa are not covered by any peptides. The hypothesis that these initial 26 aa represent a signal peptide is supported by the fact that the first two peptides mapped to the protein (starting at residue 26A) have a non-tryptic N-terminus. However, it is not the only non-tryptic endpoint observed for this peptide; for example, the peptide SIGTDTLLQIK is also non-tryptic. Below, we present an approach to distinguishing between proteolytic events and post-digestion trimming.

Non-tryptic peptides may arise from post-digestion breakup, due to hydrolysis (driven by endogenous or exogenous peptidases or by harsh chemical conditions in course of the sample preparation) or in-source decay [100]. Of the 7,079 non-tryptic peptides in the *Shewanella* dataset, 5,474 (77%) are properly contained in a longer observed tryptic peptide, and 1,605 (23%) are not. It is likely that the majority of non-tryptic peptides contained within other observed peptides result from post-digestion breakup, particularly when the longer peptide is more abundant (as estimated by spectrum count), although some of them may result from the partial proteolytic processing in vivo.

In the *S. oneidensis* project we identified 1,372 non-tryptic peptides that are not contained within any other peptide (such peptides are called *non-covered* peptides). While the 688 proteins containing non-covered peptides may be potential proteolytic targets, one can argue that these peptides may also represent (i) erroneous peptide identifications or (ii) instances where the containing tryptic peptide does not generate any observed MS/MS spectra [124]. To prove that many non-covered peptides indeed correspond to proteolytic events we point to the extremely non-uniform distribution of starting positions of these peptides along the protein (Figure 3.8). If these peptides were artifacts, we would expect to see a relatively uniform distribution of starting positions. Instead we see two pronounced peaks at positions 2 and 20, reflecting two biological phenomena: N-terminal methionine cleavage (position 2) and signal peptides (average length of signal peptides in gram-negative bacteria is $\approx$25aa [98, 104]). It should be noted that although our signal peptide peak is at 20, the distribution is skewed towards right with average signal peptide length around 26, in strong agreement with the previously reported average of 25.

To focus on these two phenomena, we limit our attention to 366 proteins in which the leftmost identified peptide is a non-covered peptide.

3.3.2 Predicting N-terminal methionine cleavage

Peptides starting at the second residue of a protein suggest cleavage of N-terminal methionine (NME) in 218 proteins. To check the effect of second residue on NME cleavage [58], we computed a cleavage efficiency factor for each of the 20 amino acids. These efficiencies of different amino acids are similar to the results observed in vitro for *E. coli* (Figure 3.9). It appears that activity in vivo may be more specific than that seen in vitro. These observations are also in close agreement with recent results independently obtained at the NIH Center for Proteolytic Pathways at Burnham Institute in the *E. coli* model using a novel labeling technology developed by Guy Salvesen at Burnham.

3.3.3 Predicting signal peptides

The average length of a signal targeting proteins to the Sec pathway in Gram-negative bacteria is 25 amino acids, with most signal peptides in the range from 20 to 30 amino acids [98, 104]. The distribution of starting positions of non-covered peptides has a pronounced peak at 20 amino acids (Figure 3.8). Of the non-covered peptides which are not explained by N-terminal methionine cleavage, 55% start at positions 21-30. Figure 3.8 suggests that most peptides in 21-30 aa window reflect signal peptides, since other 10 amino acid long windows have very few peptides. It is important to note that the bulk of protein secretion in Gram-negative bacteria occurs to the periplasmic space, therefore the corresponding processed proteins can be experimentally observed in the whole-cell extract.

We analyzed our peptide annotations in order to confirm or refute signal predictions, and possibly to discover new signal cleavage sites. In an exploratory study we examined peptides with non-tryptic N-termini and selected peptides with no upstream coverage. A clear sequence motif [28] emerges when we examine the sequence immediately upstream of these putative signal peptides predicted by MS/MS analysis (Figure 3.10). This motif closely matches motifs used by SignalP and PrediSi thus providing additional support for using non-covered peptides for signal peptide identification.

SignalP and PrediSi predict 370 and 403 proteins with signal peptides. However, there is a substantial discrepancy between these tools - only 211 signals are predicted by both tools. MS/MS evidence provides a possibility to resolve the discrepancies between SignalP and PrediSi as well as to identify signal peptides missed by both tools.

Figure 3.11a compares our predicted signal peptides with the predictions made by SignalP and PrediSi on the 1,992 expressed proteins. Our results confirm a total of 94 signal peptide predictions in vivo. In 31 cases, both SignalP and PrediSi predict signal cleavage on an expressed protein, but disagree as to the cleavage site. This ambiguity highlights the difficulties in computation signal prediction and the potential of MS/MS to confirm the correct prediction with MS/MS evidence.

On 119 of the confirmed proteins, the MS/MS results include peptides upstream of the cleavage site predicted by SignalP/PrediSi and thus represent evidence against

SignalP/PrediSi predictions. We call these the “refuted sites”. We refute 89 sites predicted by SignalP, and 38 sites predicted by PrediSi (with 8 refuted sites predicted by both tools). It is conceivable that those peptides N-terminal to the signal site come from mis-localized proteins, where the signal sequence is not cleaved. If cleavage is the norm, then peptides immediately C-terminal to the refuted sites should be seen. However, they are observed for only four of the refuted sites, and each is contained in a much more abundant (by spectrum count) fully tryptic peptide which spans the refuted site. Thus, the peptide evidence suggests that these refuted signal peptide predictions are indeed incorrect.

3.4 Conclusion

The results presented here, as well as in the subsequent chapters, demonstrate the utility of mass spectrometry in improving genomic and proteomic annotations.

Chapter 3 is, in part, a reprint of the paper “Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. 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 (2007). Genome Research. 17(9):1362-77”. The dissertation author was the primary investigator and first author of this paper jointly with Steven Tanner.

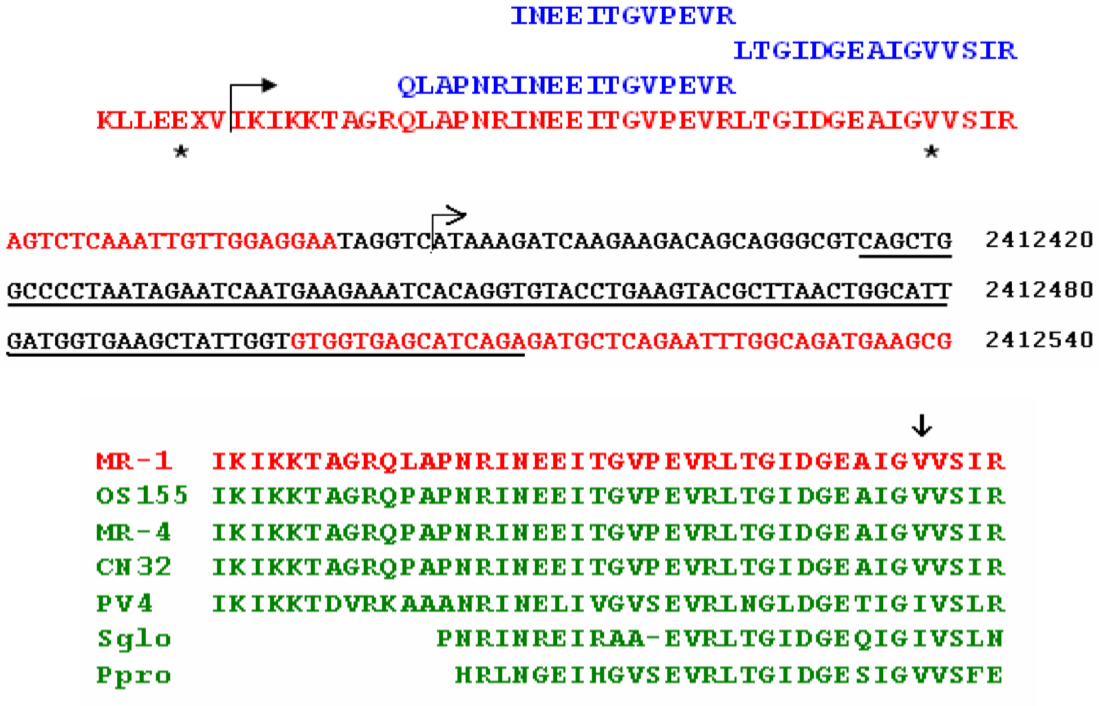
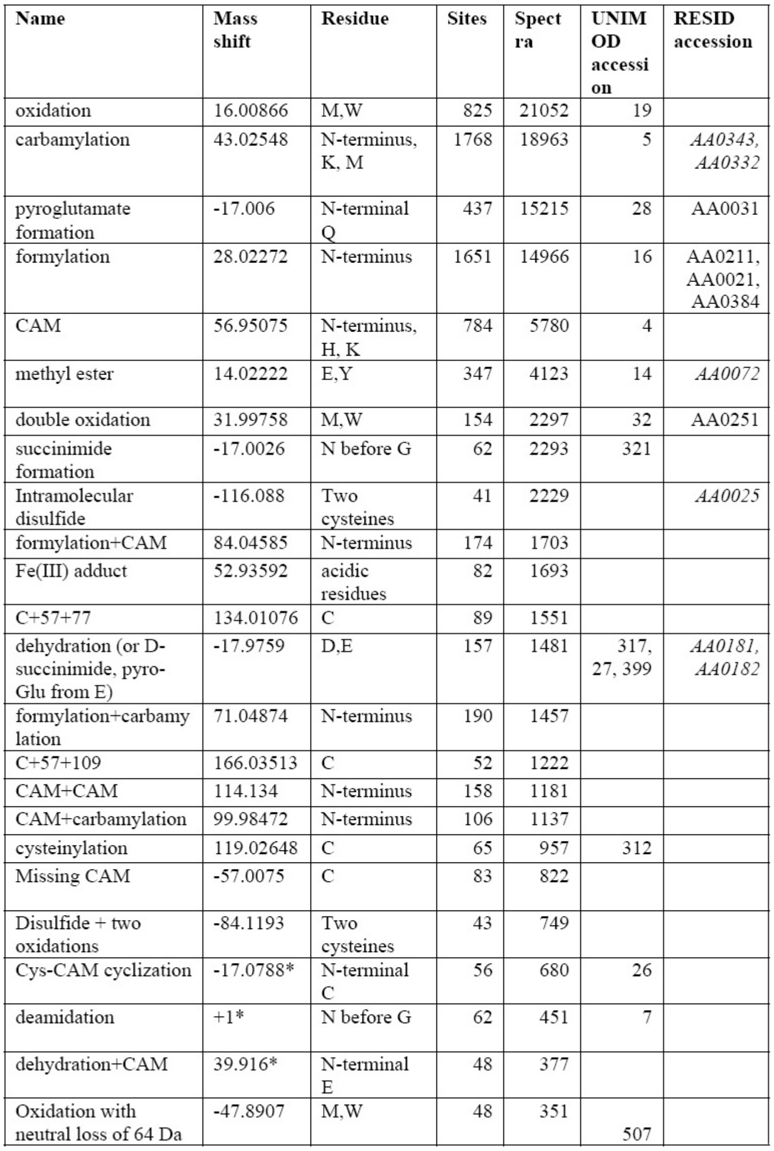
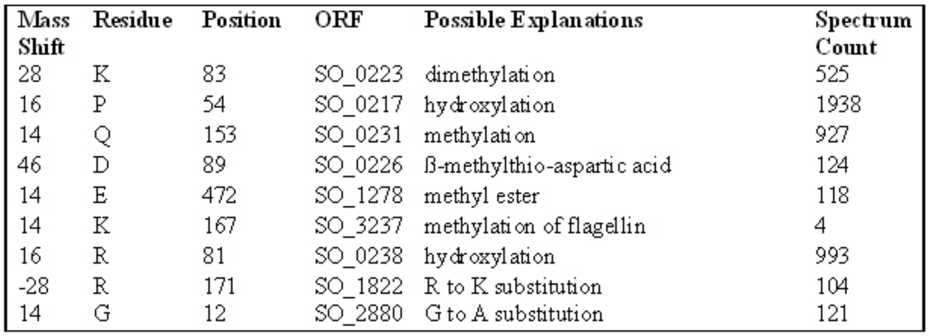

INEEITGVPEVR
LTGIDGEAIGVVSIR
QLAPNRINEEITGVPEVR
KLLEEXVIKIKKTAGRQLAPNRINEEITGVPEVRLTGIDGEAIGVVSIR
*
*
AGTCTCAAATTGTTGGAGGAATAGGTCATAAAGATCAAGAAGACAGCAGGGCGTCAGCTG 2412420
GCCCCTAATAGAATCAATGAAGAAATCACAGGTGTACCTGAAGTACGCTTAACTGGCATT 2412480
GATGGTGAAGCTATTGGTGTGGTGAGCATCAGAGATGCTCAGAATTTGGCAGATGAAGCG 2412540
MR-1 IKIKKTAGRQLAPNRINEEITGVPEVRLTGIDGEAIGVVSIR
OS155 IKIKKTAGRQPAPNRINEEITGVPEVRLTGIDGEAIGVVSIR
MR-4 IKIKKTAGRQPAPNRINEEITGVPEVRLTGIDGEAIGVVSIR
CN32 IKIKKTAGRQPAPNRINEEITGVPEVRLTGIDGEAIGVVSIR
PV4 IKIKKTDVRKAAANRINELIVGVSEVRLNGLDGETIGIVSLR
Sglo PNRINREIRAA-EVRLTGIDGEQIGIVSLN
Ppro HRLNGEIHGVSEVRLTGIDGESIGVVSFE

Figure 3.3: (a) Alignment of identified peptides (blue) with the intergenic region between TIGR proteins SO_2299 and SO_2300. The starred positions indicate the last codon of SO_2299 and the first codon of SO_2300. The arrow points to the newly postulated translational start site for SO_2300 (IF-3 gene). (b) Nucleotide sequence of the chromosome region between SO_2299 and SO_2300. The first red segment is the C-terminal end of SO_2299 and the second red segment is the N-terminal region of SO_2300. The region covered by the three identified peptides is underlined, and the arrow indicates our suggested start position for SO_2300. (c) A Tblastn comparison of the proposed new *S. oneidensis* MR-1 IF-3 N-terminus to genome sequences for *S. baltica* OS155, *Shewanella sp.* MR-4, *S. putrefaciens* CN32 , *S. loihica* PV-4, *Sodalis glossinidius* and *Photobacterium profundum* . The original start position is indicated by the arrow.



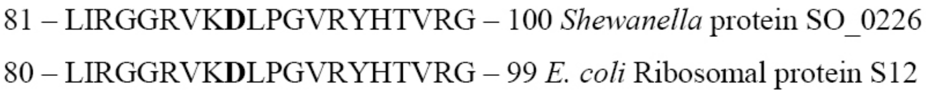
Name	Mass shift	Residue	Sites	Spectra	UNIMOD accession	RESID accession
oxidation	16.00866	M,W	825	21052	19	
carbamylation	43.02548	N-terminus, K, M	1768	18963	5	*AA0343, AA0332*
pyroglutamate formation	-17.006	N-terminal Q	437	15215	28	AA0031
formylation	28.02272	N-terminus	1651	14966	16	AA0211, AA0021, AA0384
CAM	56.95075	N-terminus, H, K	784	5780	4	
methyl ester	14.02222	E,Y	347	4123	14	*AA0072*
double oxidation	31.99758	M,W	154	2297	32	AA0251
succinimide formation	-17.0026	N before G	62	2293	321	
Intramolecular disulfide	-116.088	Two cysteines	41	2229		*AA0025*
formylation+CAM	84.04585	N-terminus	174	1703		
Fe(III) adduct	52.93592	acidic residues	82	1693		
C+57+77	134.01076	C	89	1551		
dehydration (or D-succinimide, pyro-Glu from E)	-17.9759	D,E	157	1481	317, 27, 399	*AA0181, AA0182*
formylation+carbamylation	71.04874	N-terminus	190	1457		
C+57+109	166.03513	C	52	1222		
CAM+CAM	114.134	N-terminus	158	1181		
CAM+carbamylation	99.98472	N-terminus	106	1137		
cysteinylation	119.02648	C	65	957	312	
Missing CAM	-57.0075	C	83	822		
Disulfide + two oxidations	-84.1193	Two cysteines	43	749		
Cys-CAM cyclization	-17.0788*	N-terminal C	56	680	26	
deamidation	+1*	N before G	62	451	7	
dehydration+CAM	39.916*	N-terminal E	48	377		
Oxidation with neutral loss of 64 Da	-47.8907	M,W	48	351	507	

Figure 3.4: List of common modifications (observed on at least 5 distinct sites). Carbamidomethylation (abbreviated as CAM) is added to cysteine side chains by treatment with iodoacetamide, but can be attached to other sites. Masses are computed as the average modification mass over FT spectra (except entries shown with * which had no corresponding FT spectra).



Mass Shift	Residue	Position	ORF	Possible Explanations	Spectrum Count
28	K	83	SO_0223	dimethylation	525
16	P	54	SO_0217	hydroxylation	1938
14	Q	153	SO_0231	methylation	927
46	D	89	SO_0226	ß-methylthio-aspartic acid	124
14	E	472	SO_1278	methyl ester	118
14	K	167	SO_3237	methylation of flagellin	4
16	R	81	SO_0238	hydroxylation	993
-28	R	171	SO_1822	R to K substitution	104
14	G	12	SO_2880	G to A substitution	121

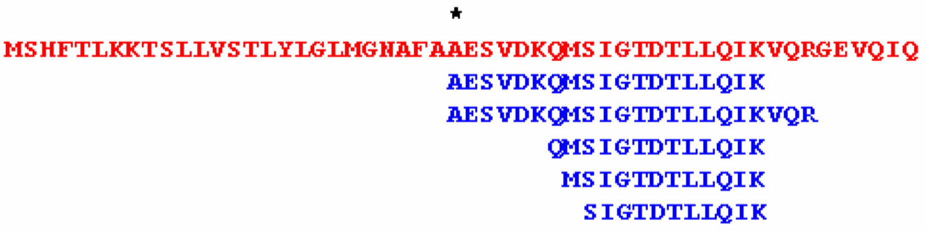
Figure 3.5: Selected PTMs supported either by studies in other bacterial genomes or by comparative genomics analysis. Some modifications are similar to previously described biological modifications in prokaryotes (typically in *E. coli*). Substitutions are supported by the presence of the target residue in other *Shewanella* strains at the orthologous residue. Hydroxylation of R on SO_0238, although not previous reported, is strongly supported by our data.



81 – LIRGGRVK**D**LPGVRYHTVRG – 100 *Shewanella* protein SO_0226

80 – LIRGGRVK**D**LPGVRYHTVRG – 99 *E. coli* Ribosomal protein S12

Figure 3.6: Position of D+46 modification on SO_0226, and its alignment with the ortholog in *E. coli*. The modified aspartate residues are shown in bold.



```
                                  *
MSHFTLKKTSLLVSTLYIGLMGNAFAAESVDKQMSIGTDTLLQIKVQRGEVQIQ
                                  AESVDKQMSIGTDTLLQIK
                                  AESVDKQMSIGTDTLLQIKVQR
                                        QMSIGTDTLLQIK
                                         MSIGTDTLLQIK
                                          SIGTDTLLQIK
```

Figure 3.7: Peptides from the N-terminal portion of conserved hypothetical protein SO_3842. The peptide breakage before the starred residue is produced when the signal peptide is cleaved and degraded. The other non-tryptic peptides are properly contained in observed tryptic peptides, and so are most likely generated by post-digestion breakup. The N-terminal ladder observed for the tryptic peptide QMSIGTDTLLQIK is a likely result of aminopeptidase-driven trimming or in-source fragmentation.

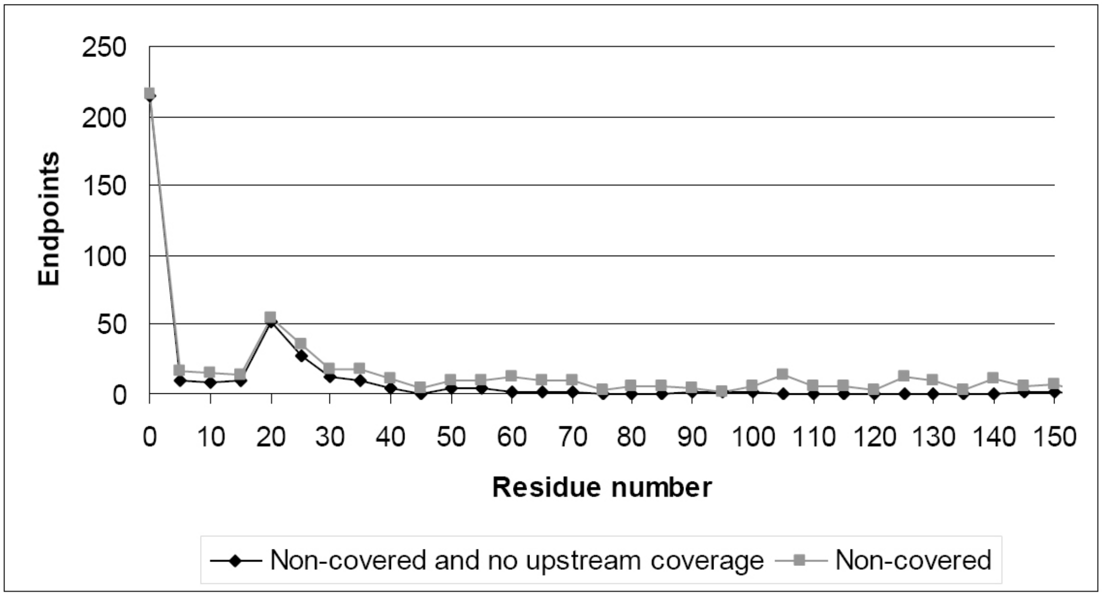


Figure 3.8: Distribution of the N-termini of all non-covered peptides, and of those which also have no upstream coverage. Two peaks observed at 2 aa and ≈20 aa. correspond to N-terminal methionine cleavage and signal peptides.

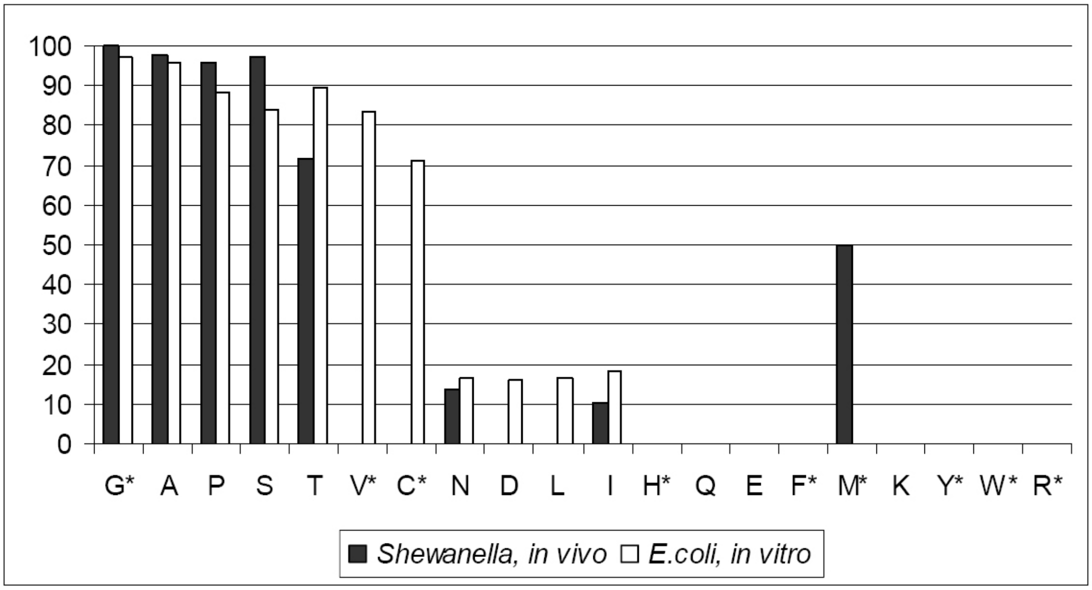


Figure 3.9: Fraction of peptides undergoing cleavage of N-terminal methionine, for a given second-position residue. Amino acids are arranged in increasing order by size of side chain. The in vitro data comes from measurements of *E. coli* MAP enzyme efficiency [58]. For the starred residues, ten or fewer N-terminal peptides were observed with that residue at the second position. In difference from [58] we observe several cases of an apparent cleavage before the second methionine in proteins starting with double methionine. However, a comparative genomics analysis of other *Shewanella* strains revealed that a large portion of them have orthologous proteins with a single methionine, rather than a double methionine. Therefore, we speculate that many proteins starting with double methionine may represent mis-annotations of the translation start site.

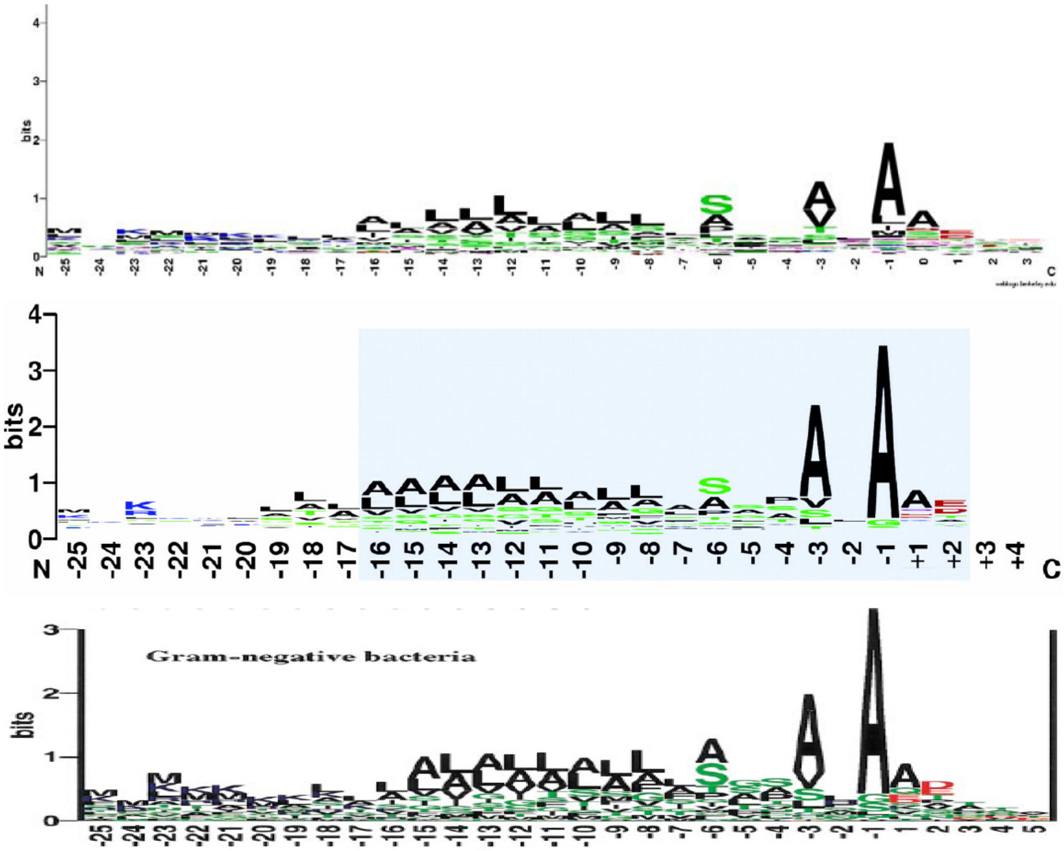


Figure 3.10: Top: Sequence logo for the amino acid sequence motif of all signal peptides identified by MS/MS analysis. Position -1 correspond to the last residue of the signal peptide. Middle: Sequence logo for gram-negative bacteria employed by PrediSi [57]. Bottom: Sequence logo for gram-negative bacteria employed by SignalP [98].

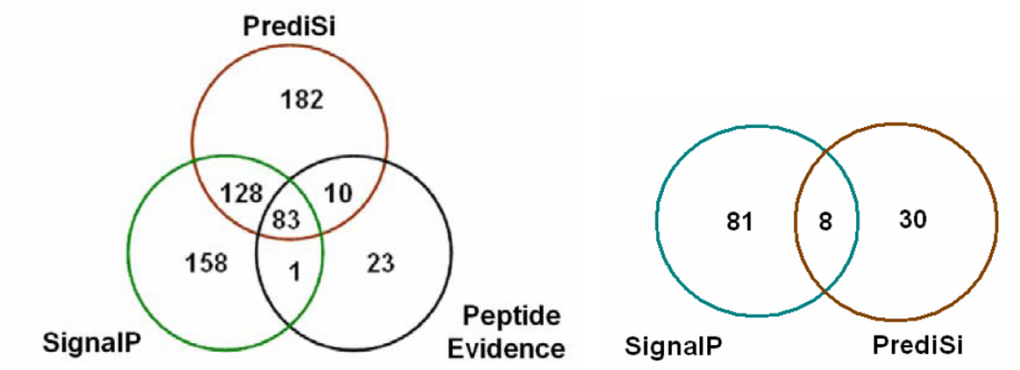
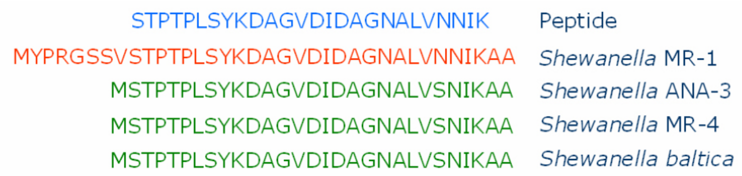


Figure 3.11: (a) Venn diagram of all signal peptide predictions on confirmed proteins. A total of 94 signal peptide cleavage sites are validated by mass spectrometry (23 of them missed by both SignalP and PrediSi). (b) Number of signal predictions by SignalP (89) and PrediSi (38) rejected due to the observation of peptides upstream of the signal cleavage site. Eight of these sites were predicted by both tools.



STPTPLSYKDAGVDIDAGNALVNNIK	Peptide
MYPRGSSVSTPTPLSYKDAGVDIDAGNALVNNIKAA	*Shewanella* MR-1
MSTPTPLSYKDAGVDIDAGNALVSNIKAA	*Shewanella* ANA-3
MSTPTPLSYKDAGVDIDAGNALVSNIKAA	*Shewanella* MR-4
MSTPTPLSYKDAGVDIDAGNALVSNIKAA	*Shewanella baltica*

Figure 3.12: Sequence of phosphoribosylformylglycinamidine cyclo-ligase (SO_2760) in *Shewanella Oneidensis* MR-1 according to TIGR annotation (red), observed non-tryptic peptide (blue) and alignment to the orthologs in other *Shewanella* strains (green).

Chapter 4

Comparative Proteogenomics

4.1 Introduction

Since the sequencing of the first genome, *H. influenzae* [44] in 1995, the number of sequenced genomes has been rising sharply. Every sequencing project is followed by annotation of the genome to identify genes, pathways, etc. Comparative genomics analysis of multiple genomes has emerged as one of the key approaches for discovery of such genomic elements that greatly improves on the existing annotation tools [11, 73, 146]. Another recent development is the application of tandem mass spectrometry (MS/MS) for genomic annotations [64, 69, 140, 42, 125, 53]. Such proteogenomic approaches further improve gene predictions and allow one to address problems that remained beyond the reach of both traditional gene prediction tools and comparative genomics.

We recently developed MS-Genome software for automated proteogenomic annotation of bacterial genomes [53] and applied it for improving annotation of *Shewanella oneidensis* MR-1, a model bacterium for studies of bioremediation and metal reduction. However, the synergy between MS/MS data from *different* species was never explored in the past. We show that such *comparative proteogenomics* analysis sheds new light on the annotations of both genomes and proteomes.

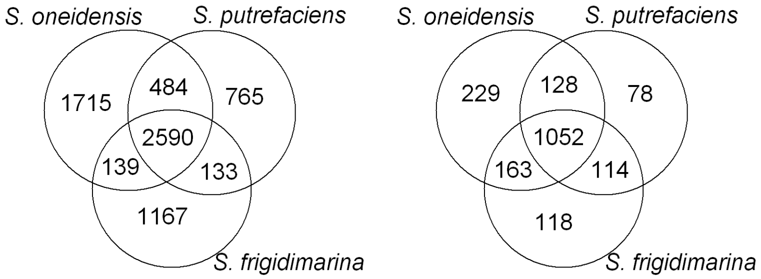
Similar to Expressed Sequence Tags (EST) studies, mass spectrometry experiments generate *Expressed Protein Tags* (EPT) that provide valuable information about expressed proteins. However, while there are 100s of studies on using ESTs for genome annotation, EPT studies are still in infancy [116]. This is unfortunate since EPTs may provide some advantages over ESTs and are easy to generate. In particular, unlike ESTs, EPTs are relatively uniformly distributed along the protein length and provide information about the translational starts, proteolytic events (e.g., signal peptides), and post-translational modifications. Also, EPTs may be less affected by splicing artifacts (like trans-splicing) and sequencing errors. However, some EPTs may represent errors in peptide identifications (and are thus completely wrong) making it non-trivial to transform the existing EST approaches into the EPT domain.

While recent high-throughput MS/MS studies generated large spectral datasets for many related species, it remains unclear how to utilize these datasets across various genomes. In this study, we analyze MS/MS datasets for three *Shewanella* bacteria representing multiple growth conditions: *Shewanella oneidensis MR-1* ($\approx$14.5 million spectra), *Shewanella frigidimarina* ($\approx$ 0.955 million spectra) and *Shewanella putrefaciens* CN-32 ($\approx$ 0.768 million spectra). These datasets provide an opportunity to analyze the *expressed* proteomes across these bacteria (henceforth referred to as So, Sf and Sp respectively). In addition to predicting new genes and finding errors in existing annotations, we show that MS/MS data helps to identify programmed frameshifts (as well as sequencing errors), a difficult problem in genomics. We demonstrate that comparative analysis of peptides across species is helpful in resolving the dilemma of "one-hit-wonders" in proteomics. We further discuss how comparative proteogenomic analysis enables identification of rare post-translational modifications and proteolytic events, two difficult problems for which the high-throughput techniques are not available. Drawing parallels from gene microarray platforms, we also use mass spectrometry based protein expression data to analyze the conserved and differentially expressed pathways across these species.

4.2 Results

4.2.1 Multiple Shewanella Genomes

The three *Shewanella* species used in this study were recently sequenced, So containing 5,131,416 base pairs being the first one [55]. Subsequently, Sf and Sp genomes have been sequenced (4,845,257 and 4,649,325 base pairs respectively). Sf



(a) (b)

Figure 4.1: Expression of orthologous genes across the three species. (a) The number of orthologs shared between different species. There are 2590 orthologous genes present in all three species (referred to as “shared genes”). (b) The number of expressed shared genes (confirmed by 2 or more peptides) among the three species . 1052 shared genes are expressed in all three species, 708 shared genes are expressed in none.

and Sp genomes, unlike So, do not have accompanying publications in the literature, although they have been cited in other studies [147]. The genome sequences and annotations used in this study were obtained from the TIGR CMR database.

The protein orthology assignments across different *Shewanella* species were prepared using Inparanoid [111], subsequently aligned by Muscle [33] (data courtesy of LeeAnn McCue and Sean Conlan). Figure 4.1(a) shows the numbers of orthologs shared by different *Shewanella* species. While 2590 genes have orthologs in all three species (we call such triplets “shared genes”), for some proteins, orthologs were found in only one other species, and in many cases (for example, 1715 in So) in none.[1]

The shared genes are used for comparative analysis in this study. The protein sequence identity between So and Sp is about 85%, while Sf is about 70% identical to the other two species (average among all shared genes). As a result, most orthologous tryptic peptides for these species differ in at least one position.

Table 4.1: Protein identification results. For each species, the total number of genes, the number of genes confirmed as expressed proteins by two or more peptides, and number of genes with only one peptide hit are reported. The numbers in the parentheses represent the number of shared genes, out of 2590 in total, that are present in the corresponding list of genes.

	S. oneidensis (So)	S. putrefaciens (Sp)	S. frigidimarina (Sf)
Annotated genes	4928	3972	4029
Expressed proteins	1967 (1572)	1625 (1372)	1744 (1447)
Single-hit proteins	404 (248)	462 (295)	464 (306)

4.2.2 Protein Identification

Based on the peptides identified from InsPecT searches (see Methods), expression of 40%-45% proteins is confirmed in each species. Table 4.1 provides the number of annotated genes and our protein identifications. Interestingly, the fraction of expressed proteins among the shared genes is much higher, at $\approx 55\%$. This hints at a correlation between protein expression and sequence conservation, in agreement with the observations made in Gupta et al., 2007. In this study, we also demonstrated the use of MS-based protein identification to analyze the expression of pathways or functional categories.

4.2.3 Resolving One-Hit-Wonders

There are 1052 shared genes that are expressed in all three species (see Figure 4.1(b)). However, in accordance with the Proteomics Publication Guidelines [21, 17], we require at least two peptides to consider a protein as expressed. Since almost every analysis of MS/MS datasets reveals a large number of proteins with a single identified peptide (one-hit-wonders), it leads to a significant reduction in the number of identified proteins (one-hit-wonders represent 21, 28 and 27 percent of all identified proteins in So, Sp and Sf respectively). For example, there are 404 such proteins in So that cannot be reported as reliable identifications. While many of them indeed represent expressed proteins, it is not clear how to separate them from erroneous peptide identifications [53].

[1] Many *Shewanella* genes may be artifacts of existing gene finding tools that tend to over-predict short genes. See [23] regarding the recent controversy on gene over-prediction.

```
Sp  -MSLLKSLAVKPLCTKLGAIAFVIAFTAGLSACAPEVGSDAWCKQMKNKPSGDWTANEAADYAKHCVFK
Sf  -MSL----------SKLFAVSSALLLTLSLTACAPEVGSEAWCKQMKEKESGDWTANEAADYAKHCVFK
So  MMFLLKLMTTKP-KVKLGAMALALAFTAGLTACAPEVGSDAWCKQMKEKPSGDWTANEAADYAKHCVFK
```

Figure 4.2: Example of correlated one-hit-wonders in shared genes. Aligned amino acid sequences of the shared gene (annotated as hypothetical lipoprotein) are shown for each organism (SO_0515 in So, CN32_3345 in Sp and Sfri_3590 in Sf). The identified peptides are shown in blue.

Below we explore the use of comparative analysis across species to reliably select the expressed proteins among the one-hit-wonders and thus remove the term "hypothetical" from some existing gene annotations.

Table 4.2: Expression signatures for shared genes. Three values in a vector correspond to three organisms, independent of the position. For example, $(0, 0, 1)$ represents shared genes that have 1 peptide in (any) one of the species, and no peptide in the other two.

Expression Signature (ES)	$(0, 0, 0)$	$(0, 0, 1)$	$(0, 0, 2)$	$(0, 1, 1)$	$(0, 1, 2)$
#proteins with given ES	434	195	182	69	187

Expression Signature (ES)	$(0, 2, 2)$	$(1, 1, 1)$	$(1, 1, 2)$	$(1, 2, 2)$	$(2, 2, 2)$
#proteins with given ES	218	10	56	187	1052

For each shared gene, we define an expression signature with three values that represent the number of peptide identifications in the three species. The value is 2 if the expression is confirmed by 2 or more peptides, 1 if only 1 peptide is observed, and 0 for no peptides. For example, the signature $(0, 1, 2)$ for a shared gene represents no peptide identification in So, 1 peptide identification in Sp and confirmed expression with 2 or more peptides in Sf. There are 27 possible distinct expression signatures that such a vector may take for a shared gene. We combine these into 10 position independent values, such that $(2, 1, 1)$ is considered the same as $(1, 1, 2)$) or $(1, 2, 1)$. Table 4.2 shows the frequency of these 10 expression signatures among the 2,590 shared genes. The argument against considering one-hit-wonders as expressed protein is that they may be unexpressed proteins with one false peptide identification. However, we note that if the orthologous genes of a one-hit-wonder are expressed in the other two species, it adds support that the gene a true expressed gene. Such genes are readily identified as having expression signature $(1, 1, 1)$, $(1, 1, 2)$ or $(1, 2, 2)$. This approach provides extra evidence for the expression of $3 \times 10 + 2 \times 56 + 187 = 329$ one-hit-wonders in total in the three species.[2]

While orthologous one-hit-wonders are strong indicators of protein expression, peptides identified at the same orthologous positions (*correlated peptides*) in different species provide an overwhelming evidence that the proteins are expressed (see Methods for description of *correlated peptides*). Since the likelihood of this happening by chance is extremely small, we now dig deeper into analysis of the orthologous one-hit-wonders and demonstrate that they often have correlated peptides. Figure 4.2 shows the example of a shared gene (annotated as hypothetical lipoprotein) that has only 1 identified peptide in each organism. However it turns out that these peptides, in spite of being slightly different from each other in their sequences, are located at the same position in the alignment of the orthologs. Thus we argue that these proteins should be considered as expressed and re-annotated to remove the term “hypothetical” from their annotations.

One reason for observing only a single peptide from a protein is the relatively few number (one in some cases) of *detectable* peptides in a protein. However, if this is the case, the orthologous peptides should be observed in the closely related species. We thus check if the only peptide observed in a protein is correlated between multiple species. If the peptide identification is spurious, it is very unlikely that the peptide will be at the same position as the observed peptides in its orthologs. Interestingly, we find 46 out of 404 one-hit-wonders in So having a correlated peptide in at least one of the other two species, providing strong evidence for the expression of these proteins. Similarly, 50 and 85 one-hit-wonders in Sf and Sp, respectively, can be resolved as expressed based on correlated peptides. We note that if the peptide identifying a one-hit-wonder is an incorrect identification, and the orthologous peptides identified in the other species are exactly the same as the one-hit-wonder peptide, they may also represent incorrect identifications of similar mass spectra (e.g., spectra from unknown contaminants). Thus, the correlated peptides are less reliable if they are identical. However, even a single change in the peptide sequences significantly changes the corresponding spectra, and

[2] The signatures (0,1.1), (0,1,2), and (0,2,2) are also useful albeit less reliable (they may represent biologically interesting cases when orthologous proteins are expressed in some species but not expressed in others).

therefore, the one-hit-wonder confirmations based on such distinct peptides are reliable. Noticeably, 38, 47 and 70 one-hit-wonders in So, Sf and Sp respectively, confirmed by correlated peptides, belong to this category.

4.2.4 Correcting Gene Predictions: Start Sites

Peptides that match the genome in the non protein-coding region upstream to a gene, within 200 bp distance, are considered candidates for early start sites. These are cases of mis-annotated genes that are shortened at their N-terminus. Cases with stop codons between the peptide and the gene start site are discarded. To avoid spurious candidates from incorrect peptide identifications, we consider a peptide only if there is another identified peptide in the same reading frame within 200 bp [53]. The starting position of the peptide (call it position X) does not necessarily correspond to the actual start site of the gene, but only tells that the actual start should be further upstream to X.

To verify early start sites and determine their exact positions, these genes were searched against proteins in 10 other *Shewanella* species, and position X for each candidate was compared to the start site of the aligned homolog. These species included *Shewanella loihica* PV-4, *S. baltica* OS155, *S. amazonensis* SB2B, *S. sp.* W3-18-1, *S. denitrificans* OS217, *S. sp.* ANA-3, *S. sp* MR-4 and *S. sp.* MR-7, besides the other two from So, Sf and Sp (leaving the one which the candidate gene belongs to). If the start site of homolog aligned to a particular position equal to or upstream of position X, then this new position was considered to be putative early start site. The most frequent (supported by maximum number of homologs) of these putative starts is chosen as the new start site for the gene.

23 among 28 such candidates in So are assigned new start sites based on the comparative analysis mentioned above. Notably, 18 of these early start sites have the expected ATG, GTG or TTG start codons, indicating that these automatically predicted start sites are indeed reliable. 2 and 3 early start sites are identified in Sp and Sf respectively.

As described in the Methods, candidates for late start sites were generated using evidence from *non-covered peptides*. Such instances indicated a potential late start site either at the beginning of the *non-covered* peptide (call it position X) or, if N-terminal cleavage occurred, one position upstream ($X - 1$). The sequences of these candidate genes are aligned to the proteins in 10 other *Shewanella* species. Each instance where the start of a protein in the other species aligns to the potential late start site (beginning at position X or $X - 1$) is considered as confirmed by comparative genomics.

In So, 5 out of 33 late start candidates are confirmed, four of which start with ATG codon and one with GTG (supporting the hypothesis that these are indeed start sites). Similarly, 11 out of 16 candidates are confirmed in Sf, and 4 among the 11 are confirmed in Sp (all of these are also found to have ATG, GTG or TTG start codon). The table also shows that the majority of these candidates have N-terminal methionine cleavage in the observed peptide. We find comparative proteomic evidence for one case where the late start site (10 amino acids downstream of the annotated start site) is conserved in the orthologs (ATP-dependent Clp protease, proteolytic subunit ClpP) between So (SO_1794), Sf (Sfri_2596) and Sp (CN32_1490). However, we note that this site is also found in our analysis of conserved proteolytic sites (below). While it is unclear whether this peptide corresponds to the late start site or a proteolytic event, it clearly represents a real non-tryptic peptide, as opposed to an incorrect identification.

We note that our approach assumes that a gene has only one translational start site. However, if there is a gene with alternative start sites, we will detect only the most upstream start site that has supporting peptide evidence.

4.2.5 Identification of programmed frameshifts and sequencing errors

A frameshift occurs when a ribosome skips one or more nucleotides in an mRNA sequence, thereby changing the reading frame to produce a different protein sequence from the original frame. In programmed frameshifts, this phenomenon is built into the translational machinery [40]. Secondary RNA structures such as pseudoknots are often responsible for the ribosomal pause and resulting frameshift [137]. While many efforts went into frameshift detection [107, 25, 19, 43, 92], accurate detection of frameshifts remains an unsolved problem. Mass spectrometry, on the other hand, provides experimental evidence for the actual translation products (proteins) and allows one to to detect the frameshifts. The presence of peptides from two different reading frames within the region of a predicted gene may represent: (1) an incorrect peptide identification, (2) an insertion/deletion sequencing error, (3) overlapping genes in different frames, or (4) a programmed frameshift. We demonstrate the application of comparative approaches for distinguishing between these possibilities.

All identified peptides are mapped to the translated frames of the genome and compared with the annotated gene coordinates to determine alternate peptide reading frames in the DNA region of a single gene. As depicted in Figure 4.3, three types of cases are typically seen. In case A, multiple peptides are observed in two different frames (only one of them being the annotated frame of the gene) in non-overlapping regions. In case B, only one peptide is observed in an alternative frame at one of the ends, while in case C, one peptide is seen out of frame with in-frame peptides on both sides. We postpone the discussion of case C since in this case incorrect peptide identifications or overlapping genes are more likely explanations than a frameshift. Case A provides the most reliable evidence of a programmed frameshift since presence of multiple peptides in the same region greatly reduces the probability that these peptide identifications are spurious. The remaining case B, with only one peptide, is ambiguous and may represent either frameshifts or incorrect peptide identifications, or overlapping genes. We exploit the sequences of multiple *Shewanella* species to find comparative evidence for putative frameshifts in these cases.

Protein sequence from the original frame of the gene, as well as sequence from the alternate frame implied by the identified peptides, is compared against the other *Shewanella* species using BLAST [2]. Good matches to the alternate-frame sequence and no matches to the gene-frame sequence provide additional evidence for a frameshift. We note that some apparent frameshifts may be caused by sequencing errors or indels in the genome sequence when a certain number (not multiple of 3) of bases are erroneously added to or deleted from the sequence. To identify such sequencing errors, we take the nucleotide sequence of the region where frameshift occurs (region between the observed in-frame and alternate-frame peptides) and generate ClustalW [22] multiple sequence alignment with the orthologous region in the other species. A sequencing error is visible in this alignment as an indel in the original sequence (see Figure 4.4). Figure 4.5 shows

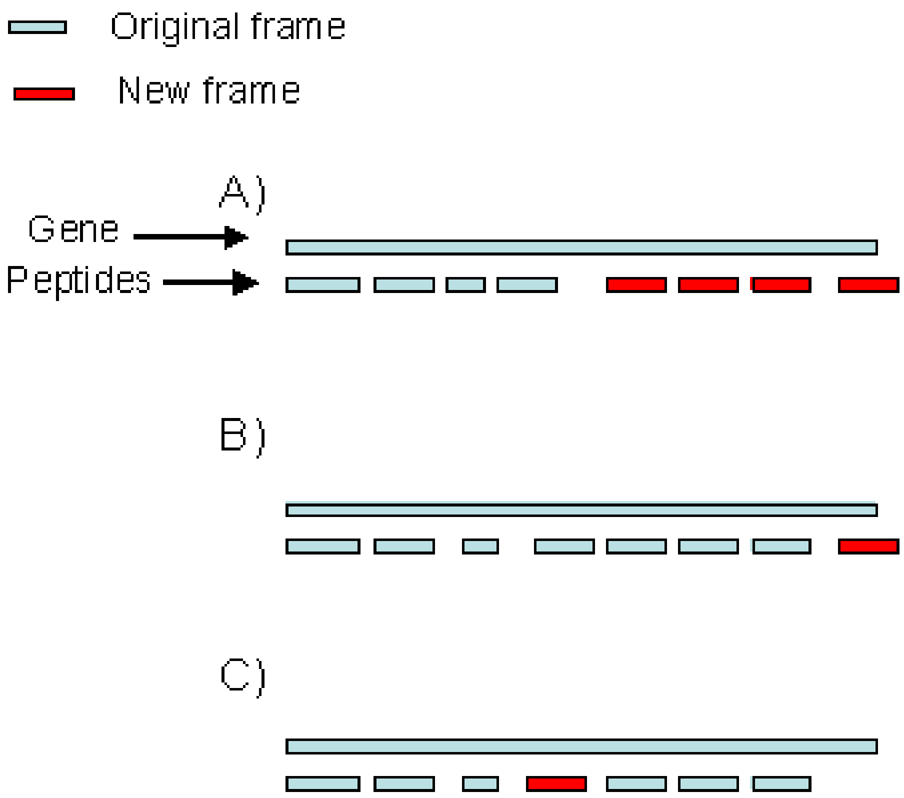
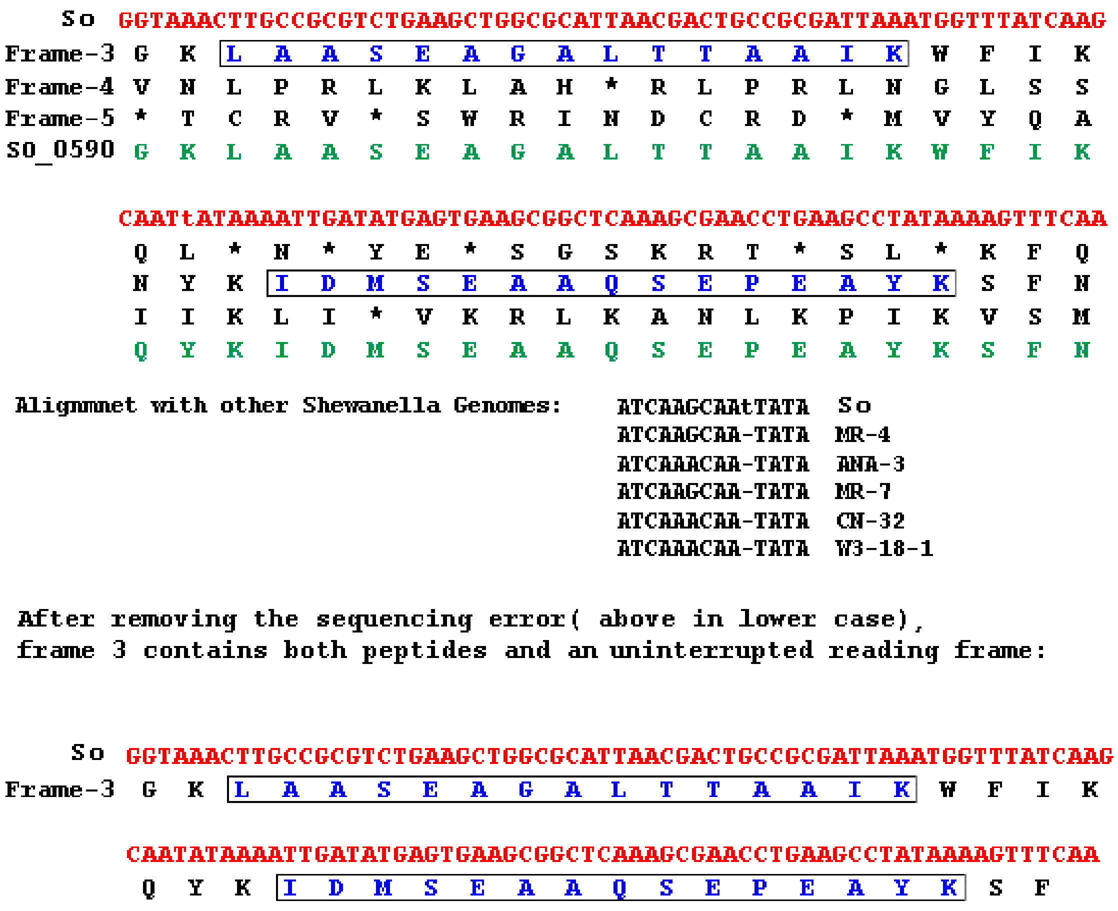
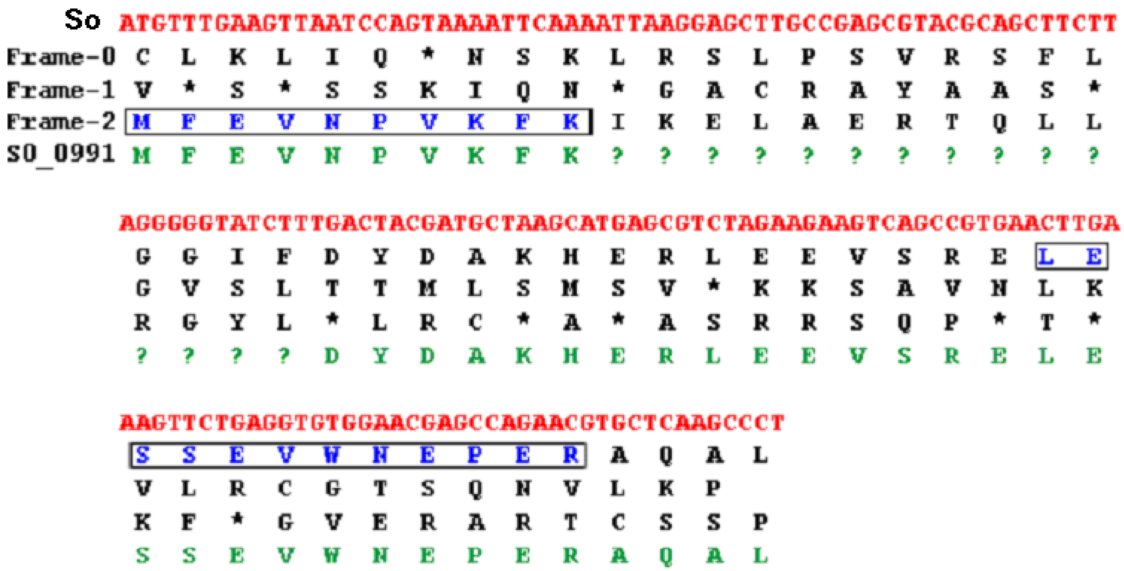


Figure 4.3: Commonly observed configurations of peptides in alternative frame. Case A: multiple peptides are observed in two different frames (one of them being the frame of the gene) in non-overlapping regions. Case B: only peptide is observed out of frame at one of the ends. Case C: one peptide is seen out of frame with in-frame peptides on both sides.



```
   So  GGTAAACTTGCCGCGTCTGAAGCTGGCGCATTAACGACTGCCGCGATTAAATGGTTTATCAAG
Frame-3 G  K  L  A  A  S  E  A  G  A  L  T  T  A  A  I  K  W  F  I  K
Frame-4 V  N  L  P  R  L  K  L  A  H  *  R  L  P  R  L  N  G  L  S  S
Frame-5 *  T  C  R  V  *  S  W  R  I  N  D  C  R  D  *  M  V  Y  Q  A
SO_0590 G  K  L  A  A  S  E  A  G  A  L  T  T  A  A  I  K  W  F  I  K

       CAATtATAAAATTGATATGAGTGAAGCGGCTCAAAGCGAACCTGAAGCCTATAAAAGTTTCAA
        Q  L  *  N  *  Y  E  *  S  G  S  K  R  T  *  S  L  *  K  F  Q
        N  Y  K  I  D  M  S  E  A  A  Q  S  E  P  E  A  Y  K  S  F  N
        I  I  K  L  I  *  V  K  R  L  K  A  N  L  K  P  I  K  V  S  M
        Q  Y  K  I  D  M  S  E  A  A  Q  S  E  P  E  A  Y  K  S  F  N

 Alignmnet with other Shewanella Genomes:     ATCAAGCAAtTATA  So
                                              ATCAAGCAA-TATA  MR-4
                                              ATCAAACAA-TATA  ANA-3
                                              ATCAAGCAA-TATA  MR-7
                                              ATCAAACAA-TATA  CN-32
                                              ATCAAACAA-TATA  W3-18-1

 After removing the sequencing error( above in lower case),
 frame 3 contains both peptides and an uninterrupted reading frame:

   So  GGTAAACTTGCCGCGTCTGAAGCTGGCGCATTAACGACTGCCGCGATTAAATGGTTTATCAAG
Frame-3 G  K  L  A  A  S  E  A  G  A  L  T  T  A  A  I  K  W  F  I  K

       CAATATAAAATTGATATGAGTGAAGCGGCTCAAAGCGAACCTGAAGCCTATAAAAGTTTCAA
        Q  Y  K  I  D  M  S  E  A  A  Q  S  E  P  E  A  Y  K  S  F
```

Figure 4.4: Frameshift generated by sequencing error. In top panel, the nucleotide sequence for gene SO0590 is shown in red, the amino acid sequence of the protein is shown in green, and the amino acid sequences of the three translated frames are shown in black. Peptides identified by mass spectrometry are marked in blue (surrounded by boxes). The middle panel shows the ClustalW alignment with other *Shewanella* species in the region where frameshift occurs. The erroneous insertion of an extra “t” stands out in the alignment. The bottom panel indicates that both peptides fall in the original frame if the extra nucleotide is removed.



```
     So ATGTTTGAAGTTAATCCAGTAAAATTCAAAATTAAGGAGCTTGCCGAGCGTACGCAGCTTCTT
Frame-0 C  L  K  L  I  Q  *  N  S  K  L  R  S  L  P  S  V  R  S  F  L
Frame-1 V  *  S  *  S  S  K  I  Q  N  *  G  A  C  R  A  Y  A  A  S  *
Frame-2 M  F  E  V  N  P  V  K  F  K  I  K  E  L  A  E  R  T  Q  L  L
SO_0991 M  F  E  V  N  P  V  K  F  K  ?  ?  ?  ?  ?  ?  ?  ?  ?  ?  ?

        AGGGGGTATCTTTGACTACGATGCTAAGCATGAGCGTCTAGAAGAAGTCAGCCGTGAACTTGA
        G  G  I  F  D  Y  D  A  K  H  E  R  L  E  E  V  S  R  E  L  E
        G  V  S  L  T  T  M  L  S  M  S  V  *  K  K  S  A  V  N  L  K
        R  G  Y  L  *  L  R  C  *  A  *  A  S  R  R  S  Q  P  *  T  *
        ?  ?  ?  ?  D  Y  D  A  K  H  E  R  L  E  E  V  S  R  E  L  E

        AAGTTCTGAGGTGTGGAACGAGCCAGAACGTGCTCAAGCCCT
        S  S  E  V  W  N  E  P  E  R  A  Q  A  L
        V  L  R  C  G  T  S  Q  N  V  L  K  P
        K  F  *  G  V  E  R  A  R  T  C  S  S  P
        S  S  E  V  W  N  E  P  E  R  A  Q  A  L
```

Figure 4.5: An example of a programmed frameshift. The nucleotide sequence for gene SO_0991 is shown in red, the amino acid sequence of the corresponding protein is shown in green, and the amino acid sequences of the three translated frames are shown in black. This gene has been correctly annotated in TIGR, and our predicted peptides in both the original frame and the alternative frame match the protein sequence.

an example of a programmed frameshift detected through this approach.

We identified 12 frameshift candidates in So conforming to case A.All these candidate frameshifts were verified with significant E-values. Nine of these instances are estimated to be sequencing errors, and three genes are putative programmed frameshifts: SO0991 (+1), SO4538 (-1), and SO4115 (-1). SO0991 (Figure 4.5) is related to the peptide chain release factor 2 in *E. coli*, that is known to undergo a programmed frameshift [27]. 15 frameshift candidates were identified conforming to case B but not verified by comparative evidence. No frameshifts candidates could be verified in Sp or Sf. This may be attributed to the relatively small number of spectra for these two species (less than a million spectra each) as compared to 14.5 million spectra for So.

4.2.6 Proteolytic events

In Gupta et al., 2007, we demonstrated the use of genome scale MS/MS dataset for identification of N-terminal proteolytic events such as N-terminal methionine cleav-

```
           ↓
KRIESFGSQVYLVNTGWTGGPHGIGKRFDIPTT

 RIESFGSQVYLVNTGWTGGPHGIGKR  (4)
 RIESFGSQVYLVNTGWTGGPHGIGK  (10)
  IESFGSQVYLVNTGWTGGPHGIGK  (6)
           LVNTGWTGGPHGIGK  (98)
           LVNTGWTGGPH  (1)
            VNTGWTGGPHGIGK  (11)
             NTGWTGGPHGIGK  (9)
```

Figure 4.6: A cleavage site located within a peptide ladder. The first line shows a section of the protein SO_0162(residues 399-432) with the cleavage site between Y and L marked by a downward arrow. The subsequent lines show the identified peptides along with their spectral counts in the parentheses.

age and signal peptide cleavage. An in vivo proteolytic event can be observed as a non-tryptic peptide (assuming the proteolytic enzyme does not have the same specificity as trypsin). However, non-tryptic peptides may also be observed due to other reasons, such as degradation of tryptic peptides or incorrect peptide identifications. In Rodriguez et al., 2008, we showed that the likelihood of incorrect peptide identifications can be reduced drastically (to less than 0.1%) by considering only *doubly-confirmed* cleavages and filtering out possible degradation products [112].

By applying the same filtering approach as in Rodriguez et al., 2008 and removing the cuts explained by the trypsin specificity, we obtain 365, 130 and 62 putative proteolytic sites in So, Sp and Sf respectively. To check whether some of these sites are conserved between multiple organisms, we map them on the alignment of orthologous protein. 31 proteolytic sites are found conserved between two or more organisms (see Table 4.3). This is a significantly larger number of conserved sites than expected by chance. For example, with proteomes of length $\approx$ 1 million amino acids each, the expected number of sites conserved by chance between Sp and Sf is less than $(62/10^6) \times (130/10^6) \times 10^6 \approx 0.01$, but we observe 13. One may further challenge that these cleavages may be an artifact of in-vitro peptide degradations, and that these peptides may be over-represented in proteins containing multiple peptides. In this case, the statistical argument above must to be applied to the set of these highly-expressed proteins rather than to all proteins. To check this, we took proteins with 10 or more peptides (635 proteins in Sp, 671 in Sf) with total length close to 300 thousand aa in each organism, and 128 and 57 putative proteolytic sites in Sp and Sf respectively. All 13 sites conserved between Sp and Sf belong to these highly-expressed proteins. The expected number of sites conserved by chance in these proteins is $(128/300000) \times (57/300000) \times 300000 \approx 0.02$, still much smaller than the observed 13 sites. Thus we argue that the conserved sites reported here cannot be results of non-specific degradations.

We note that many of these sites are located within peptide ladders (multiple overlapping peptides) which also raises the possibility that these cleavage sites may be a result of peptide degradation (see example in Figure 4.6). However, carefully looking into these ladders, we see that they are more likely a union of two peptide ladders, one coming from the proteolysed and the other from the unproteolysed protein product. This is supported by high spectral counts for the peptides around the cleavage site in many cases, given that one expects much lower spectral counts (usually 1) for degraded peptides as compared to the tryptic (un-degraded) peptide in a ladder. For example, the peptide LVNTGWTGGPHGIGK that supports the predicted cleavage site in Figure 4.6 has a spectral count of 98, even higher than the spectral counts of the covering tryptic peptides. Based on this and the statistical evidence shown above, we expect that our conserved cleavage sites represent in-vivo proteolytic events. Since the knowledge of proteolytic events in bacteria is still very limited at genomic scale, we are not able to provide additional supporting information about the origin or relevance of each predicted site individually; but we make the data available for comparison with future studies.

Note that here we used the traditional rules for trypsin specificity, allowing a cut after arginine or lysine but not before proline. Interestingly, 5 of the 31 conserved sites happen to be cuts between arginine and proline, indicating that these may be a result of trypsin digestion, further supporting the conclusion in Rodriguez et al., 2008 that the cuts after arginine and lysine followed by a proline should be considered tryptic. Other 7 sites are signal peptide cleavages also predicted by SignalP [13] providing additional

Table 4.3: List of conserved proteolytic sites. The first column indicates the number of organisms in which the site was observed. The next three columns tell the name of the protein containing the site and the position (in parentheses) of the cleavage site within the protein. The last column indicates if the site is actually a cut between arginine and proline (denoted by R.P), or a signal peptide cleavage site.

# organisms	Protein in So	Protein in Sp	Protein in Sf	Comment
2	SO3420(20)	CN32_2738(20)		Signal
2	SO0162(409)	CN32_3571(409)		
2		CN32_2230(328)	Sfri_2257(328)	R.P
2	SO2402(20)	CN32_2042(20)		
3	SO0231(196)	CN32_3759(196)	Sfri_0148(196)	
2	SO2328(14)	CN32_1875(14)		
2	SO0234(255)		Sfri_0151(255)	
2	SO0235(58)	CN32_3755(58)		
2		CN32_3753(212)	Sfri_0154(212)	
2		CN32_3750(37)	Sfri_0157(37)	
2	SO2746(19)		Sfri_1464(19)	Signal
2		CN32_1517(28)	Sfri_2626(28)	
2	SO1816(21)	CN32_1510(21)		Signal
2		CN32_1495(281)	Sfri_2585(281)	
3	SO1794(9)	CN32_1490(9)	Sfri_2596(10)	
3	SO1638(23)	CN32_1357(20)	Sfri_1279(20)	Signal
2		CN32_1348(47)	Sfri_1270(47)	
2	SO1351(202)	CN32_1162(202)		
3	SO3649(204)	CN32_0981(204)	Sfri_3087(204)	R.P
3	SO0992(210)	CN32_3049(210)	Sfri_0583(210)	R.P
3	SO0951(21)	CN32_0891(21)	Sfri_0664(30)	Signal
2	SO0929(349)		Sfri_0646(349)	R.P
2	SO0781(286)	CN32_3209(286)		
2	SO4078(247)	CN32_0594(247)		
2	SO4509(52)	CN32_0337(52)		
2	SO0424(870)	CN32_3417(870)		
2		CN32_3415(149)	Sfri_3775(149)	R.P
2	SO0432(363)	CN32_3409(363)		
2	SO0432(235)	CN32_3409(235)		
2	SO0610(18)	CN32_3274(18)		Signal
2	SO3904(23)		Sfri_3332(23)	Signal

support that our detected sites represent proteolytic events rather than statistical artifacts.

4.2.7 Post-translational modifications

In this section, we use the term post-translational modification (PTM) to denote chemical modifications of individual residues, such as phosphorylation, oxidation, methylation etc.[3] Blind PTM searches with MS-Alignment [136] or ModifiComb [117] find all possible mass offsets (revealing potential modifications) without *a priori* knowledge of which modifications may be present in the sample. The first applications of these tools revealed that the world of modifications is much larger than previously thought [144,99] and, at the same time, emphasized the still unsolved problem of finding rare modifications. Since blind searches may yield thousands of modifications [53], the "strength in numbers" approach [136] considers frequent modifications (e.g., offset +16 on M) as reliable and discards rare modifications as unreliable. A comparative version of this approach would be to identify modifications that are seen in multiple samples. After the post-processing of MS-Alignment results as described in Methods, we find 162 distinct modifications that are observed in all three species. While 74 of these represent chemical adducts that are expected in mass spectrometry experiments, 88 others reveal biologically interesting modifications as well as other potentially important modifications that remain unknown.

The "strength in numbers" approach, while successful, leaves many rare modifications unexplained. These modifications may either represent rare and biologically important modifications, or incorrect peptide identifications. However, it is very unlikely to find a modification at the same site in orthologous genes in two different species just by chance (especially if the peptides are not identical). We find 48 such modifications that are conserved at one or more sites in the genome. For example, 48 on W is found to be conserved at three different sites. At two of these sites, the peptides covering the orthologous modification position are not identical, virtually eliminating the possibility of incorrect identifications. Most of these modifications are previously unknown, providing a refined set of candidates for experimental validations.[4] While post-translational

[3] Mass spectrometry experiments reveal both *in vivo* and *in vitro* modifications (chemical adducts).

[4] Experimental validation of these modification requires chemical synthesis and remains beyond the

modifications must be important in the metal-reducing *Shewanella* species, studies of modifications in *Shewanella* are still in infancy [128]. Although there are currently no reported experimental studies that can be used for verification of our comparative proteogenomic predictions, we hope that our analysis provide sufficient evidence to warrant some experimental verifications. Note that we cannot claim the biological significance of identified modifications; they could be either in-vivo PTMs or in-vitro chemical adducts, although the low-frequency modifications are less likely to be conserved if they are introduced in-vitro after digestions.[5]

4.3 Methods

4.3.1 Peptide Identification

Peptide identification in So was described in the earlier study [53]. The MS/MS spectra were acquired on ion-trap mass spectrometers (LCQ, ThermoFinnigan, San Jose, CA) using electrospray ionization. We use InsPecT [126] (July, 2007 version) to search the spectra of each species against a database containing the six-frame translation of the genome along with common contaminants and a decoy database of the same size. Inspect search was run using default parameter settings (fragment ion tolerance of 0.5 Da and parent mass tolerance of 2.5 Da). The InsPecT score threshold is selected for each case to limit the number of identifications on the decoy database to at most 1% of the number of identifications on the target database, to keep the false discovery rate under control. After the filtering step, we obtained 29,160 peptides in So, 22,820 peptides in Sf and 22,358 peptides in Sp. These include 337, 222 and 269 peptides in So, Sf, Sp respectively that do not match the annotated proteins in these genomes. We demonstrate that coordinated mapping of these peptides (that are usually discarded as false identifications) represents valuable information for improving genome annotations.

scope of this study.

[5] We also cannot exclude the possibility that they represent a "combined" modification, i.e. two different modifications (let's say with offsets X and Y) on neighboring residues that is misidentified as a single modification (with offset X+Y). However, many of our identifications have excellent b/y ladders indicating that such artifacts are unlikely.

4.3.2 Analyzing late start codons

We describe an algorithm for predicting “late” start codons, i.e., the (correct) start codons that are located *downstream* of the wrongly annotated start codons. While a late start codon implies a “missing” peptide in the beginning of the protein (between the wrongly annotated and correct start codons), such missing peptides can also be caused by low peptide detectability [79] or may simply represent signal peptides. However, *non-covered* peptides (non-tryptic peptides with no upstream coverage, see [53] for more details) in the beginning of the protein, that cannot be explained by the signal peptide consensus sequence, point to late start codons. There are 33 cases of N-terminal-most non-covered peptides in So, within 18 residues of the start. Conspicuously, many of them either begin with ATG start codons or start immediately after a start codon (as in the case of N-terminal Methionine cleavage, see [53]). If all these peptides were artifacts, the distribution of the codons for amino acids at positions 1 (where the observed peptide begins) and -1 (corresponding to N-terminal Methionine cleavage) in these peptides would be somewhat uniform with average $33/61 \approx 0.5$ peptides per codon. Instead, we see a non-uniform distribution at position 1 and -1 with a sharp peak at ATG (standard Methionine start codon) and over-representation of other start codons (TTG and GTG). We thus believe that all these cases cannot be artifacts (such as degradation products or incorrect peptide identifications).

To exclude signal peptides from consideration, we consider only non-covered peptides located within a distance of 18aa or less from the start of the protein (signal peptides are typically longer than 18 aa). 33, 16 and 11 candidates are observed in So, Sf and Sp respectively. Comparative analysis of the three *Shewanella* species is subsequently performed to validate these candidates for late start codons.

4.3.3 Correlated peptides

Traditional MS/MS analysis is focused on identification of proteins and is less concerned with the question of which peptides in a protein are observed or not observed. In this study, we utilize the availability of proteomic data from related species to analyze the expression of peptides at orthologous positions. In a typical mass spectrometry experiment, some peptides with low detectability are always missed, resulting in a highly non-uniform protein coverage by identified peptides [109, 79]. For example, while most ribosomal proteins in So have high coverage (above 50%), a few have low coverage and one of them does not have any identified peptides. Peptide detectability may depend on several factors including protein abundance, peptide length, peptide hydrophobicity etc. and several groups are using large datasets to develop the ability to its prediction [124, 88, 86].

All identified peptides in shared genes were mapped to the alignment of the orthologs to get their coordinates with respect to the alignment. This provides a uniform reference scale to compare the positions of observed peptides between the orthologous proteins in the three species, as individual proteins may have different lengths. Peptides identified by MS/MS in two species are called *correlated peptides* if they are observed in the same position in the protein alignment or one of them *spans* another. In other words, if one peptide is located at positions $(start_1, end_1)$ in the alignment, and the other peptide at $(start_2, end_2)$, then peptides are considered *correlated* if $start_1 \leq start_2 \leq end_2 \leq end_1$ or $start_2 \leq start_1 \leq end_1 \leq end_2$.

4.4 Identification of post-translational modifications

MS-Alignment [136] was used to identify PTMs in each of the three organisms in a blind mode, in the range from -200 to +250 Daltons. Common contaminants like keratins were included in the protein sequence databases. A decoy database of the same size as the actual protein database, containing shuffled sequences, was used to control the error rates. Any hits to the decoy database are expected to be incorrect identifications. A score cut-off is chosen such that the number of PTM sites identified in the decoy database is at most 5% of the number of identifications in the target database. This provides a controlled PTM-site specific false-discovery rate of 5%. We note that this is a more stringent criterion than a 5% error rate at the spectrum or peptide level, since several peptides in the forward database may point to the same PTM-site. We further removed all spectra that were identified in the regular Inspect search. After this post-processing of MS-Alignment results, 9917, 7649 and 6709 PTMs were obtained in So, Sf and Sp respectively. We only use tryptic modified peptides in the subsequent analysis.

4.5 Discussion

Shewanella oneidensis MR-1 is among the most carefully annotated bacterial genomes: gene predictions in this genome were studied in two papers [95, 29] and are being continuously improved by the Shewanella Federation http://www.shewanella.org. Significant manual effort (that took into account comparative genomics evidence) also went into the annotation of *Shewanella frigidimarina* and *Shewanella putrefaciens* CN-32. We demonstrate that comparative proteogenomics approach leads to improved annotations even for these well-studied genomes, let alone for genomes with only automated annotations available. Recent proliferation of low-cost DNA sequencing techniques will soon lead to an explosive growth in the number of sequenced genomes and will turn manual annotations into a luxury that can be afforded for only a small fraction of newly sequenced genomes. We therefore suggest that complementing DNA sequencing projects by comparative proteogenomics projects can be a viable alternative approach to improve both genomic and proteomic annotations.

Chapter 4 is, in part, a reprint of the paper “Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes. 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 (2008). Genome Research. 18:1133-1142”. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Does trypsin cut before proline?

5.1 Introduction

Trypsin is arguably the most commonly used enzyme in mass spectrometry based proteomics to digest proteins into peptides. One of the reasons for trypsin's success in mass spectrometry is that it cuts exclusively after arginine and lysine [100]. However, the rules for trypsin specificity (sometimes referred to as "Keil rules" [71]) are rather involved and have long been defined in terms of the amino acids on either side of a potential cut site. The commonly accepted rule for a trypsin cut site is [RK].[ˆP] i.e. [RK] at position before the cut (P1 position) and [ˆP] at positions after the cut (P1' position) forms a trypsin cleavage site. In this *regular expression* notation, [RK] denotes "either R or K", and [ˆP] denotes "any amino acid other than P". Similarly, it is also believed that trypsin activity is suppressed if acidic residues are present on either side of the cleavage site or in the case of cysteine at C-terminal. These rules are part of the standard descriptions of commercially available trypsin from leading vendors like Sigma-Aldrich and Promega and description of trypsin specificity at the popular ExPASy (Expert Protein Analysis System) web server [http://expasy.org]. Leading peptide identification tools like Mascot [105], X!Tandem [26], or ProteinProspector [24] incorporate [RK].[ˆP] rule for trypsin specificity into their search algorithms.

Using these rules as a filtering criteria, these algorithms expect to remove false hits and, at the same time, make the searches faster by reducing the search space. Since the number of identified peptides is affected by these criteria, it is important to have accurate rules for trypsin specificity when applied to peptide identification. Below we demonstrate that enforcing the Keil rule in MS/MS database search is actually counter-productive since it leads to losing a significant number of peptide identifications. In fact, we found that the number of peptides produced by supposedly “illegitimate” [RK].[P] cleavages is higher than the number of peptides produced by legitimate [RK].[C] cleavages and comparable to the number of peptides produced by [RK].[W] cleavages. We therefore argue that for all practical purposes, the MS/MS search engines should either remove [RK].[ˆP] filtering rule or should complement it with [RK].[ˆC] and [RK].[ˆW] rules. We further observe that while some previously formulated rules describing subtle variations in trypsin specificity (for example, inhibition of cleavage after K in CKY, DKD, CKH, CKD, and KKR as described at ExPASy server [http://www.expasy.ch/tools/peptide-mass-doc.html] are supported by MS/MS data, others (inhibition of cleavage after R in RRR, CRK, DRD, RRF, KRR) are not.

From the pragmatic perspective, the improved description of trypsin specificity leads to only a modest increase in the number of identified peptides in typical MS/MS database searches. More importantly, the careful description of trypsin specificity is crucial for emerging *labeling-free* approaches to studies of proteolysis. While the existing simplistic view of trypsin specificity has been acceptable for standard MS/MS searches, it becomes inadequate for more complex analysis, such as using MS/MS for studies of regulatory proteolysis [131,37]. In such studies, the sample is digested with both trypsin and a regulatory protease (e.g., a caspase) with the goal to identify the specificity of the regulatory protease. MS/MS is then employed to determine all cleavages in the resulting sample. Afterwards, one has to “subtract” trypsin cleavages from all found cleavages to find the cleavages made by the regulatory protease. However, if the model of trypsin specificity is inaccurate, these studies are likely to fail. Another area that will require detailed knowledge of trypsin specificity is the analysis of *in-vivo* proteolytic events in the sample subjected to trypsin with the goal to infer the natural proteolytic cleavages induced by various proteases (without attempting to infer the specificities of individual proteases). We recently demonstrated that many proteolytic events can be derived from *labeling free* MS/MS shotgun data [53]. For example our analysis of N-terminal methionine excision (NME) without any labeling in [53], showed excellent correlation with NME cleavages revealed by labeling approaches [131] (Guy Salvesen, personal communication). Extending these labeling free approaches to other proteolytic events requires a very accurate description of native trypsin specificity.

Several computational and experimental methods have been used previously to derive protease specificities [70, 130, 110, 54, 138]. In [110], positional scanning synthetic peptide combinatorial libraries were used to identify the specificity of interleukin-1 converting enzyme. A similar approach, using combinatorial fluorogenic substrate libraries, was developed in [54]. Turk et al. [138] used pooled sequencing of peptide library mixtures to determine specificity of six matrix-metalloproteases. Recently, attempts have been made to utilize recombinant technology based methods [31], using phage display to expose vast number of recombinant peptides to a given protease and amplifying the peptides released through specific cleavage. More recently, Boulware et al., 2006 [16] developed whole-cell protease activitiy assay by displaying fluorescent reporter substrates on the surface of *E. coli* as N-terminus fusions. Our study complements these approaches by capitalizing on large MS/MS datasets that bypass the need to generate combinatorial libraries.

The Keil rules for trypsin specificity were derived two decades ago based on *in vitro* analysis of a relatively small sample of substrates. The follow up studies [129] were limited to a few hundred amino acid tetramers to derive the trypsin specificity rules. Large MS/MS datasets present an opportunity to increase the number of experimentally confirmed cleavage sites by two-three orders of magnitude and to re-examine the Keil rules on much larger samples obtained *in vivo*. However, such analysis requires caution since erroneous peptide identifications and peptides resulting from post-digestion breakup may give the false impression of proteolytic cleavages. Recent developments in mass spectrometry instrumentation have enabled large scale experiments, generating millions of spectra [141, 53, 1]. These datasets provide unprecedented opportunity to study the specificity of the protease used for protein digestion. Given a large peptide list (obtained without any assumptions about the enzyme specificity), we demonstrate that it is possible to derive the specificity rules *de-novo*.

In this study, we specifically question the commonly accepted rule that trypsin does not cut after R or K if the C-terminal residue is P. We first show that a substantial number of peptides are identified that have a cut between [RK] and P, with a strict control on the rate of false identifications. We further limit ourselves to *doubly confirmed cleavages* to keep the error rate extremely low. We consider several possibilities that may result in [RK].P cuts, whether they are results of post-digestion breakup, trypsin digestion or digestion by a contaminant protease. We evaluate statistical evidence for these hypotheses and all feasible alternatives indicate that cleavages before proline represent real peptides (as opposed to false peptide identifications). We thus argue that [RK].P cleavages should be "legitimatized" in peptide identification algorithms.

Here we have suggested modifications to rules of trypsin specificity, for application in mass-spectrometry, for one specific case: with proline at P1 position. We further suggest that an exhaustive study is required to reconsider all possible modifications to trypsin and other enzymes' specificity rules, in more general fashion allowing for longer motifs. Only recently we have begun to generate large scale datasets that can allow for such empirical determination of these rules, complementing our existing knowledge of enzyme specificities.

5.2 Methods

5.2.1 Doubly-Confirmed Cleavages

28,377 peptides were identified from the *Shewanella oneidensis* MS/MS dataset containing 14.5 million spectra. We look at the two endpoints of all identified peptides to analyze trypsin specificity rules. However, we note that as many as 5% of our peptide identifications may be incorrect, thereby yielding nearly 2800 incorrect endpoints. To avoid any bias in the specificity rules introduced by such errors, we limit analysis to doubly-confirmed cleavages, i.e. cleavages that are supported by endpoints of at least two peptides (see Figure 5.1). The figure shows different configurations of peptides that may lead to a doubly confirmed cleavages.

Assuming that incorrect identifications are randomly distributed in the database, the expected number of endpoints that are shared by two incorrect peptides is extremely small. In this random distribution model of incorrect identifications, we note that the probability of seeing two adjacent peptides (sharing their C-terminal and N-terminal endpoints respectively) is the same as the probability of seeing two overlapping peptides both sharing their C-terminal (or both N-terminal) endpoints, and in either case, the shared endpoint is reliably identified. Thus we obtain a set of reliable 11085 peptide endpoints, that are shared between 21661 peptides (some peptides may have both their termini contributing to the endpoints). On the contrary, the corresponding number of doubly confirmed cleavages in the reverse database was found to be only 11, three orders of magnitude lower than in the actual database (false discovery rate of less than 0.1%). Thus, nearly all doubly confirmed cleavages represent real cleavages.

5.2.2 Post-digestion breakup

A critical challenge in using mass spectrometry derived peptides for analyzing protease specificity lies in differentiating between the peptides produced by the protease digestion and those produced by post-digestion breakup (degradation) of other peptides. While the former type of peptides are representative of the proteases that were present at the time of digestion, the post-digestion breakup products are expected to be somewhat random in their endpoints. To minimize any bias introduced by such breakup products, we filter out any doubly confirmed endpoints that are likely to be formed by degraded peptides. Trimming of one or two amino acids from a peptide is a commonly observed degradation pattern [53]. Cases III and V in Figure 5.1 represent examples of degraded peptides that may lead to a doubly confirmed endpoint, and should be removed. Accordingly, we filter out a doubly confirmed endpoint if there exists another peptide that extends beyond the endpoint by one or two amino-acids. For example, in case III of Figure 5.1, the C-terminus of peptide QMSIVSYGEEK extends beyond the doubly confirmed endpoint E.E by two amino acids (EK), and is thus filtered out. This filtering step removes 1294 endpoints.

5.3 Results

We re-examine the rules for trypsin specificity by looking at the di-amino-acid pair (di-AA) between which the peptide endpoints are located. The frequency of each

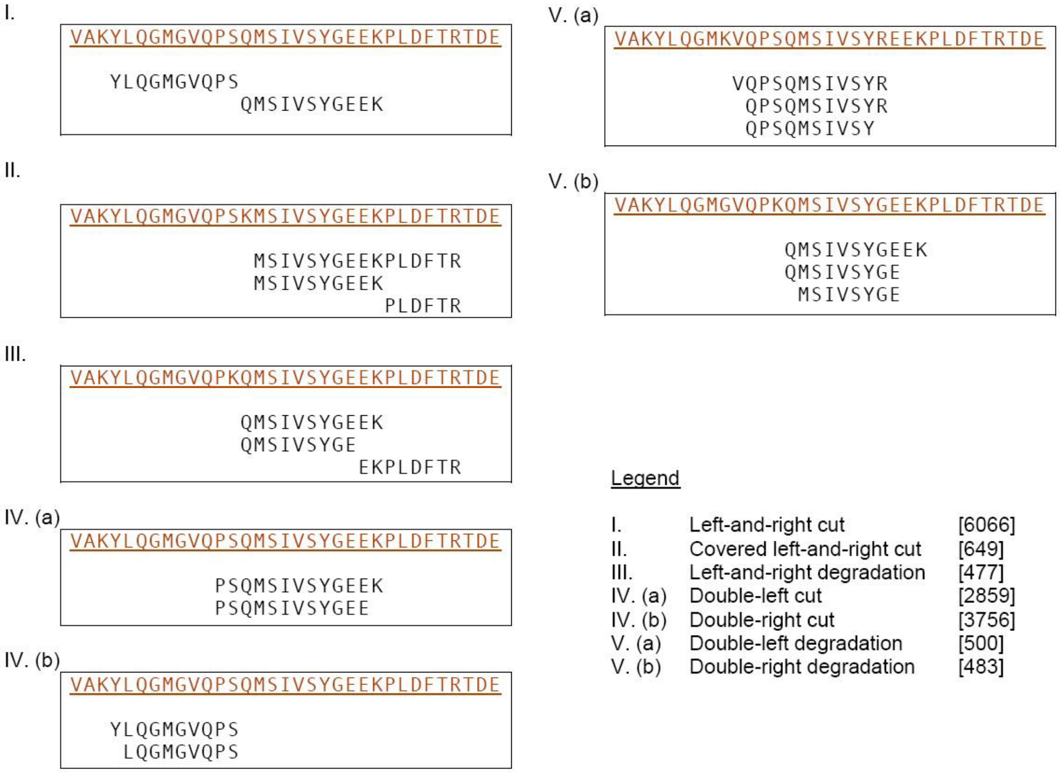

I.
VAKYLQGMGVQPSQMSIVSYGEEKPLDFTRTDE
YLQGMGVQPS
QMSIVSYGEEK
II.
VAKYLQGMGVQPSKMSIVSYGEEKPLDFTRTDE
MSIVSYGEEKPLDFTR
MSIVSYGEEK
PLDFTR
III.
VAKYLQGMGVQPKQMSIVSYGEEKPLDFTRTDE
QMSIVSYGEEK
QMSIVSYGE
EKPLDFTR
IV. (a)
VAKYLQGMGVQPSQMSIVSYGEEKPLDFTRTDE
PSQMSIVSYGEEK
PSQMSIVSYGEE
IV. (b)
VAKYLQGMGVQPSQMSIVSYGEEKPLDFTRTDE
YLQGMGVQPS
LQGMGVQPS
V. (a)
VAKYLQGMKVQPSQMSIVSYREEKPLDFTRTDE
VQPSQMSIVSYR
QPSQMSIVSYR
QPSQMSIVSY
V. (b)
VAKYLQGMGVQPKQMSIVSYGEEKPLDFTRTDE
QMSIVSYGEEK
QMSIVSYGE
MSIVSYGE
Legend
I. Left-and-right cut [6066]
II. Covered left-and-right cut [649]
III. Left-and-right degradation [477]
IV. (a) Double-left cut [2859]
IV. (b) Double-right cut [3756]
V. (a) Double-left degradation [500]
V. (b) Double-right degradation [483]

Figure 5.1: Different types of doubly confirmed cleavages. The legend tells the type of the endpoint: left indicates that the concerned endpoint is at the N-terminus of the observed peptide, and right indicates the C-terminus. Numbers in the brackets indicate the frequency of each type observed in our dataset.

di-AA is counted, and the sorted list (in decreasing order) is shown for the top 45 di-AAs in Figure 5.2(a). This list supports Olsen et al., 2004 [100] observation that trypsin only cuts after R and K since the only di-AA deviating from this rule is A.A. This is readily explained by the presence of many *Shewanella oneidensis* signal peptides with cleavage motif AXA.A [53]. Other less frequent cleavages (like L.A and N.S) can potentially be explained by signal peptide with weaker recognition sites (like A.K or L.A) or still unknown proteolytic events in *Shewanella* (like N.S).

Of the 400 possible pairs, if the Keil rule for trypsin specificity is correct, we expect the list to be dominated by 38 [RK].[ˆP] pairs. However, from the pragmatic perspective (maximizing the number of peptide identifications), Figure 5.2(a) demonstrates that favoring [RK].[^P] over [RK].[^C] does not make sense statistically. Moreover, the number of [RK].[W] cleavages is not dramatically larger than the number of [RK].[P] cleavages. These rankings may be affected by the relative occurrence of different di-AAs in the proteome, hence we normalize the observed di-AA frequency with its frequency in the proteome (Figure 5.2(b)). The normalized frequency of a di-AA is same as its raw frequency multiplied by 100 and divided by its frequency of occurrence in the confirmed proteins. Surprisingly, R.P and K.P, which are considered non-tryptic sites, are seen ahead of all other non-tryptic sites (with the exception of A.A). This is very unlikely to happen by chance if these di-AAs are non-tryptic like all other di-AAs, and rather suggests that R.P and K.P are tryptic but perhaps with less propensity for cleavage than other [RK].* sites. Therefore, we argue that since [RK].C and [RK].W are recognized as legitimate trypsin cuts, [RK].P should also be considered valid.

There are three plausible explanations for observing [RK].P cleavages: (A) The cut between [RK] and P is a result of the post-digestion breakup of an originally tryptic peptide; (B) The cut between [RK] and P is induced by trypsin and (C) induced by other contaminant protease.

Since we filter out obvious post-digestion breakup products (as described in Methods) and still get many [RK].P cleavages, chances of (A) being the dominant source of such cleavages is low. However, one argument in favor of (A) is that the amide bond N-terminal to proline can be hydrolyzed upon collision induced dissociation MS/MS or by nozzle-skimmer fragmentation [100, 61]. Depending on whether this breakup is seen only after arginine or lysine can affect our conclusions, therefore we consider two sub-hypothesis separately: (A1) The breakup before proline is seen irrespective of the amino acid N-terminal to it; (A2) the breakup before proline is seen only when [RK] is at N-terminal. If (A1) is true, the expected ratio of [RK].P and [^RK].P type of cuts is roughly 2 to 18 (or slightly different if we adjust to the differences in frequency of the 20 amino acids in the proteome). However the observed number of doubly confirmed [RK].P cuts observed in our dataset is 31, while the number of [^RK].P cuts is only 8. This clearly suggests that (A1) can not be the case, and if (A) is true then (A2) must be true, since (A1) and (A2) are mutually exclusive.

Sequencing grade modified trypsin (Promega), as described in [91] was used in

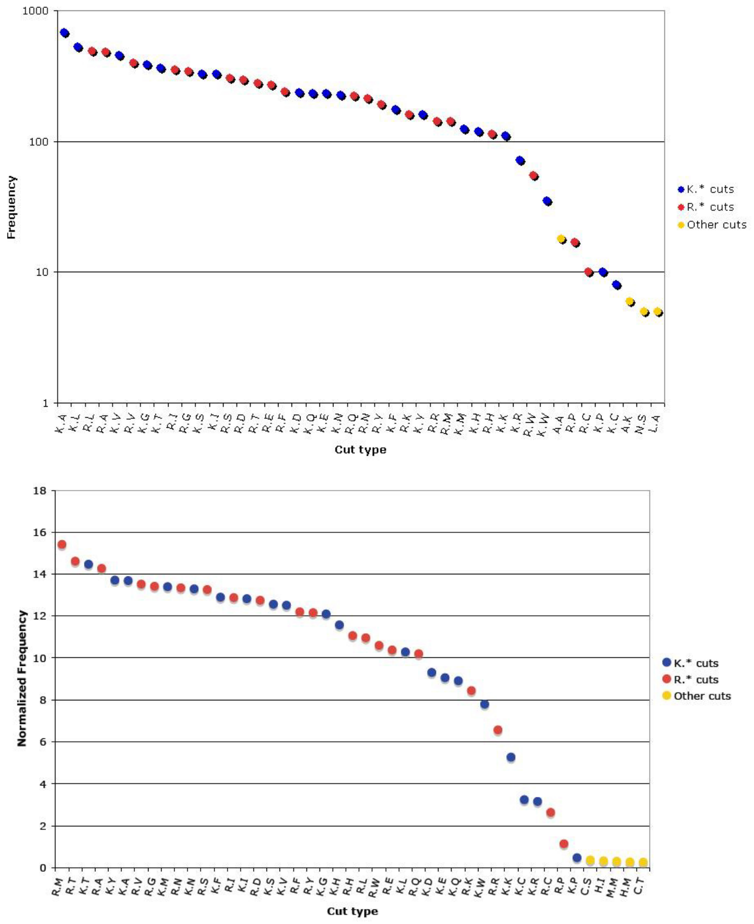


Figure 5.2: (a) Raw and (b) Normalized frequencies of 45 most frequent di-AA pairs. Y-axis in (a) is drawn on logarithmic scale. In (b), the normalized frequency of a di-AA is same as its raw frequency multiplied by 100 and divided by its frequency of occurrence in the confirmed proteins. While this figure is largely consistent with a previous observation [10] that the cleavage rate is higher after arginine than after lysine, it also shows three deviations from this rule: K.Y, K.F, and K.H have higher cleavage efficiency than R.Y, R.F, and R.H. We remark that our analysis did not support the previously suggested rule [71] that trypsin cuts before proline only in the triplets WK.P and MR.P. We observed no WK.P and MR.P cleavages but a surprisingly large proportion of AK.P and AR.P cleavages.

the experiments for digestion and standard laboratory protocols were followed (the same setup has been used in multiple studies at the Pacific Northwest National Laboratories and other places). The possibility of contamination by other proteases, either in the commercially available trypsin or during the experimental procedures, cannot be completely excluded and this might be a source of [RK].P cleavages instead of trypsin actually, suggesting hypothesis (C). However, since the protocols used here are fairly standard, the same cleavage patterns are likely to be seen in other experiments using trypsin for digestion, even if some cleavages are not induced by trypsin itself. Thus we argue that for practical applications, if there are any differences in activity of 100% pure trypsin compared to the activity of commonly used trypsin samples in mass spectrometry, the specificity rules based on the latter should be used in the analysis.

While we do not have sufficient evidence in this study to resolve between hypotheses (A2), (B) and (C), we argue that they all have the same practical implications. Each of these cases indicates that [RK].P cuts are legitimate cleavages corresponding to real peptides. Whether these are preferred breakups between [RK] and P (A2), trypsin induced cleavages (B) or common contaminant protease induced (C), they represent valid peptides like other [RK].* cuts, and should not be disregarded by the peptide identification algorithms. Either the trypsin specificity rules used in these algorithms should be altered or separate rules should be added allowing [RK].P cuts. For the sake of simplicity, we recommend the former alternative.

5.4 Discussion

Surprisingly, despite the importance of tryptic digests, there are very few MS/MS-based rigorous statistical studies of trypsin specificity. This may be caused by the prevalent and deceivingly simple “after R or K unless followed by P” rule, which is a fairly good representation of trypsin specificity for most applications. As a result, most such studies were limited to verification of this rule rather than attempting to challenge or modify the rule. In this work, we assume no prior knowledge of trypsin specificity and attempt to rigorously derive it by analyzing the cleavage patterns (inferred from large MS/MS datasets) in “discovery” rather than in “validation” mode. Our results imply that the “after R or K unless followed by P” rule is often violated. We emphasize that our analysis refers to commercially available trypsin commonly used in MS/MS experiments. While these products may have contaminants that introduce non-tryptic cleavages to the sample, it is unlikely that they specifically introduce lysine-proline and arginine-proline cleavages that we believe represent trypsin activities. We realize that the trypsin specificity may vary between different commercial vendors and between different techniques for sample preparation. Our proposal to abandon the Keil rule refers to typical vendors and to typical sample preparations rather than to an “ideally” purified trypsin or “perfect” sample preparations.

Our approach is not limited to trypsin and can be used for analyzing specificity of various proteases used in preparing samples for MS/MS experiments (the specificity of some of them is poorly understood). A number of recently emerged MS/MS approaches are based on “protease cocktails” that add a variety of proteases (like Lys-C) to the standard tryptic digestion with the goal to improve protein identifications [145, 87, 7, 9]. These new developments often require knowledge of specificity of the proteases commonly used in protease cocktails, and the approach described here can be helpful for such studies.

Most MS/MS database search programs have a parameter that limits the number of uncleaved sites. Since the cleavage before Proline is not favored but can happen, we suggest that the search programs allow cleavages before Proline but not count them as an uncleaved sites. Likewise, our data suggest that it may also make sense either not to count the K.K and K.R cuts or to weigh them differently.

Finally, we remark that in this work we did not explore the use of the elution time to analyze the post-digestion breakup (alternative A). If the peptides arise after LC separation (from dissociation in the MS interface), the parent should be observable in the same parent spectrum. Since two peptides will typically have greatly different elution times it might be possible to rule out the alternative A.

Chapter 5 is, in part, a reprint of the paper “Does trypsin cut before Proline? J. Rodriguez, N. Gupta, R.D. Smith and P.A. Pevzner (2008). Journal of Proteome Research. 7(1):300-5”. The dissertation author was the second author and investigator of this paper.

Chapter 6

Improving detection of proteolytic sites

6.1 Introduction

This study extends the label-free approach developed in Chapter 5 to analyze the specificity of three other proteases used for digestion in mass spectrometry. Using multiple enzymes for digestion can be helpful for increasing the peptide-coverage of proteins, or in applications where overlapping peptides are desirable, such as in the construction of spectral networks and de novo protein sequencing [9, 8]. *Staphylococcus aureus* V8 protease (also known as Glu-C), chymotrypsin and CNBr are popular alternatives to trypsin. Here we empirically derive the known specificity rules for these proteases and present evidence for some notable deviations from these rules, suggesting that the reaction specificity is not as simple as previously assumed.

While Rodriguez et al. 2008 [112] introduced the doubly-confirmed cleavages to infer reliable cleavage sites, we illustrate that it is possible to determine equally reliable but significantly larger list of cleavage sites using MS-GeneratingFunction [75]. We show that comparative analysis of multiple digests allows one to reliably identify N-terminal methionine excisions, signal peptide cleavages and other putative proteolytic events using our MS-Proteolysis software tool. MS-Proteolysis can be used to analyze any MS/MS dataset (including ones that were not generated to study proteolysis) to discover in vivo proteolytic events.

6.2 Results

6.2.1 Peptide Identification

High-throughput LC-MS/MS experiments (see Methods) generated 1.51 million, 1.24 million and 1.54 million spectra for *Shewanella oneidensis* MR-1 sample digested with V8 protease, chymotrypsin and CNBr respectively. These spectra were analyzed with InsPecT [126] using the default settings (fragment ion tolerance of 0.5Da and parent mass tolerance of 2.5Da). *Shewanella oneidensis* MR-1 protein sequences obtained from TIGR Comprehensive Microbial Resource, were used as the protein database (total size ≈1.5MB). A decoy database of the same size (containing shuffled protein sequences) was used to estimate the peptide-level False Discovery Rate (FDR) and limit it to 5% (the spectrum-level FDR is less than 2%). 31630, 9390 and 5317 peptides were identified in V8 protease, chymotrypsin and CNBr digests respectively.

Using MS-GeneratingFunction [75] at the stringent 0.1% FDR, 19868, 6388 and 3442 peptides were identified in V8 protease, chymotrypsin and CNBr digests respectively. We also analyzed the previously published *Shewanella* samples digested with trypsin [53] with MS-GeneratingFunction and identified 32531 peptides at 0.1% FDR.

MS/MS spectra from trypsin digests of *Saccharomyces cerevisiae* proteome were obtained from the PeptideAtlas repository [32], and analyzed using InsPecT and MS-GeneratingFunction as in the case of *Shewanella*. 7488 peptides were identified at 0.1% FDR.

6.2.2 Reliable cleavage sites

Each identified peptide reveals two cleavage sites through its end-points. A *doubly-confirmed* cleavage site is defined as a position in the proteome which is an end-point for two or more identified peptides [112]. 8635, 2146 and 866 such sites were identified for V8 protease, chymotrypsin and CNBr digests respectively. To ensure that the peptides considered in this analysis are produced by the protease and not by post-digestion breakup, a filtering step is applied before constructing the final list of doubly-confirmed cleavages [112]. In Rodriguez et al., 2008 [112], the error rate for doubly-confirmed sites was found to be only 0.1% when the peptide level error rate was

5%, as in this study. Therefore, less than 9 among all doubly-confirmed cleavages in V8 protease, 2 in chymotrypsin and 1 in CNBr are expected to be false positive identifications.

A protease substrate is conventionally labeled as ..P5, P4, P3, P2, P1, P1', P2', P3', P4', P5'.., where the cleavage is between P1 and P1' positions. The commonly used specificity rules for the three proteases studied here are based on the amino acid at P1 position. V8 protease is known to cleave after acidic residues D and E, CNBr is known to cleave after M and chymotrypsin is known to cleave after aromatic amino acids Y, F and W and partially after L. To compare these rules with the cleavages observed in our dataset, we analyze the fraction of different amino acids at the P1 position (Table 6.1). Figure 6.1 illustrates that the amino acids expected at the P1 position by known specificity rules are indeed highly over-represented at that position in our identified cleavages, thus supporting the rules as well as showing that our mass spectrometry-based approach can independently derive the specificity rules without prior knowledge. We now focus on the disagreements between the two to see if the observed cleavages can be used to refine the known specificity rules for these proteases.

To extend the analysis from just P1 position to a longer motif around the cleavage site, we constructed the sequence logos [28] for regions containing P15 to P15' positions, shown in Figure 6.2. The figure indicates that P1 position indeed plays the dominant role in determining the specificity of all three proteases. P1' position reveals a small signal, which might represent a secondary preference contingent upon P1 position. To analyze this, we categorize each cleavage site by the pair of amino acids (di-AA) between which the site is located (P1 and P1' positions). The observed frequency distribution for di-AAs flanking the cleavage sites can be used to better infer the specificity of the protease used for digestion [112]. Since not all di-AAs are equally likely to occur in the proteome, we normalize their observed frequencies by their background amino acid frequencies in *Shewanella* proteome. In the following three sections, we use this data to analyze specificity rules for each protease in detail. We will use the notation X.Y to represent a di-AA, where X is the amino acid at P1 position and Y is the amino acid at P1' position (use of * for X or Y indicates that *any* amino acid can be present at that position).

6.2.3 V8 protease specificity

V8 protease is expected to cleave after D and E [59, 122], as is also observed in our data (Figure 6.1). The figure also shows that cleavages after E are more likely than cleavages after D, in agreement with previous observation [5]. Austen et al., 1976 [5] claimed that the protease does not cleave between E and P, while such cleavages were supported by Houmard et al., 1972 [59]. We find that E.P di-AA has rank 33 among all di-AAs, well ahead of many D.* cleavages like D.Q and D.R. In fact, the relative frequency of E.P cut is similar to the relative frequency of E.Q cut (rank 27). This suggests that V8 protease does cleave between E and P, although the propensity of such cleavages is lower than other E.* sites. We also notice very low propensity of E.E, D.D, and D.E cleavages suggesting that in such cases V8 cleaves after the second amino acid. In contrast, however, E.D cleavage is frequent.

While the standard rule suggests that the top 40 di-AA cleavage sites should be D.* and E.*, followed by a random mix of other di-AAs, we surprisingly find 7 G.* cleavages (G.A, G.S, G.M, G.H, G.T, G.G, G.N) among the top 50. This leads to a new hypothesis that V8 protease also cleaves after G, although less efficiently than after D and E. Cleavages after G have not been previously reported for this protease. We constructed the sequence logo to look at the sequence patterns around cleavage sites that have G at P1 position. Figure 6.3 shows the sequences logo for these sites, and for comparison, the logo for sites that have E at the P1 position. While the sites with E at P1 show only modest preferences at P1' position and none at other positions, the sites with G at P1 position show a larger motif involving P2, P3, P1' and P2' positions. For example, F and Y are over-represented at P2 position while A is over-represented at P3, P1' and P2'. Relatively lower preference for G, as compared to D or E, at P1 position may indicate that a longer sequence motif is needed for these sites to be recognized by the protease. To ensure that these trends observed for G are specific to cleavages and do not reflect a general preference of G to co-occur with certain amino acids, we constructed the sequence logo for all positions containing G in whole *Shewanella* proteome. Figure 6.3c shows that there is no such bias in the proteome; therefore, the patterns observed here are specific to the cleavages.

6.2.4 CNBr specificity

CNBr is known to cleave after M [48], and this is also clearly visible in the observed frequency table (Table 6.1). It appears that cleavages are less likely when the amino acid at P1’ position is Q or T. No cleavages are observed between M and M, which indicates that in such cases of adjacent possible cleavage sites, CNBr cleaves at the second site. We also do not see any cleavages between M and W, and between M and C; however, this may be because of the low frequency of these di-AAs in the proteome.

While cleavages with M at P1 position are predominant in the observed list of di-AA pairs, we find that CNBr also shows a minor preference for R and K at P1 position (Figure 6.1). In fact, among the ranks 15 to 55 of top di-AA pairs for this protease, 16 have K at the P1 position while 17 have R at the P1 position (while only 2 of each type were expected by chance). This suggests that besides its primary specificity, CNBr may also have a small propensity to cleave after the basic amino acids. Table 6.1 shows that while R and K have some preference to be at P1 position in all proteases, the trend is particularly strong for CNBr indicating the role of protease in these cleavages.

6.2.5 Chymotrypsin specificity

Figure 6.1 indicates that chymotrypsin cuts after a number of different amino acids. Chymotrypsin is usually expected to cleave after F, Y and W [118]. However, this rule is not unanimous, and some studies also include L in this list [119]. While F and Y stand out at P1 position, we observe that the preference for H, K, L, M and R at P1 is comparable to the preference for W (after adjusting for background distribution of amino acids). Constructing a sequence logo for these unexpected sites (Figure 6.4(b)) indicates that positions P3, P2, P1’ and P2’ are relatively more important for specificity at these sites than for the expected sites with F, Y or W (Figure 6.4(a)). Alanine is found to be the most commonly present amino acid at these positions among the unexpected sites, indicating its possible role in determining the specificity. While we cannot totally discard the possibility of trypsin contamination in these samples, we argue that if such contamination is commonplace, it is practical to include R and K in the *empirical* specificity rules of chymotrypsin (and subtract the cleavages with R/K at P1), especially when using mass spectrometry for discovery of in vivo proteolytic events.

6.2.6 Using MS-GeneratingFunction for identifying cleavage sites

When using mass spectrometry-derived peptides to infer the specificity of proteases, it is critical to have extremely low error rates at the peptide level. This becomes even more important when the goal is to analyze the secondary (lower frequency) cleavage preferences, to minimize the possibility that erroneous peptide identifications are mis-attributed as secondary cleavage sites. The notion of doubly-confirmed cleavage sites allowed us to limit the analysis to very reliable cleavage sites [112]. However, this approach may be too restrictive, since we do not always expect multiple peptides to begin or end at all real cleavage sites in the proteome (particularly for low-abundance proteins).

Below we suggest that the same level of stringency can be achieved with even higher sensitivity, if one can control the False Positive Rate (FPR) of *individual* Peptide-Spectrum matches, without restricting to doubly-confirmed cuts. Recently, Kim et al., 2008 [75] described MS-GeneratingFunction approach that computes the FPR of individual peptide identifications, as opposed to the False Discovery Rate of *all* peptide identifications computed using the standard target-decoy approaches. MS-GF, therefore, not only controls the overall error rate but also ensures that every individual peptide identification selected above the threshold is reliable. In particular, it identifies a much larger number of peptides with virtually 0% FDR (i.e, no peptides identified in decoy database for the same score threshold) than other popular tools (see [75]).

We used MS-GeneratingFunction to calculate the FPR of peptide identifications for each of the three proteases, and thresholds were chosen to limit the FDR to 0.1% (at par with doubly-confirmed cuts). From each identified peptide, a cleavage was inferred at its N-terminus. We noticed that cleavages at C-termini show increased frequency of basic amino acids (R/K) at P1 position (perhaps due to ionization bias and/or detection preferences of existing MS/MS database search tools) and, therefore, were not included in the current analysis[1]. Note that similar over-representation of carboxy-terminal R

[1] Note that trypsin contamination alone does not explain this bias, since in that case, even the cleavages at N-termini of the peptides are expected to show similar preference for R/K at their P1 position, even if

and K residues in peptide identifications has been previously reported in native peptides even in absence of trypsin digestions [89].

Possible post-digestion breakup products of intact peptides were filtered off as described earlier [112]. 13116, 5116 and 2698 cleavage sites were detected in V8 protease, chymotrypsin and CNBr digests respectively, a significant increase compared to the doubly-confirmed cleavage approach.

Thus, MS-GeneratingFunction can be used to detect reliable cleavage sites with the same stringency and accuracy as doubly-confirmed cleavages. The larger list of cleavage sites obtained through this approach, however, can be particularly valuable for detecting in vivo proteolytic events, as discussed in the next section.

6.2.7 Detection of putative in vivo regulatory proteolytic sites

While most of the cleavages detected in the proteome (after discarding the post-digestion breakup products) are generally expected to be produced by the protease used for digestion, biological samples may also contain some in vivo cleavages representing N-terminal methionine excisions, removal of signal peptides, and other regulatory proteolytic events.By subtracting the cleavage sites explained by the specificity of the protease (e.g., cleavages after R and K for trypsin), one can filter the list of all cleavage sites to find candidates for such in vivo proteolytic events [53]. However, extra evidence is usually required to confirm these candidate sites as regulatory proteolytic sites. For example, Gupta et al., 2008 [51] compared candidate sites from three *Shewanella* species to find a set of evolutionary conserved putative proteolytic sites. Here, we argue that detection of a cleavage site across different digests of the same proteome can also provide evidence to confirm in vivo proteolytic sites.

From the list of cleavage sites obtained by MS-GeneratingFunction at 0.1% error rate, the sites explained by the protease specificity were excluded. The specificity rules were kept broad (based on the results obtained above) to minimize the possibility of any in vitro cleavage being considerd as an in vivo cleavage. We excluded all cleavages with the following amino acids at P1 position: D/E/G for V8 protease, M/R/K for CNBr, and F/Y/W/L/R/K/M/H for chymotrypsin. Besides the three proteases analyzed in this

those R/K residues are not present in the detected peptides.

study, we also used trypsin as the fourth protease for increasing the coverage, using data from Gupta et al., 2007 [53] (sites with R/K at P1 position were excluded [100]). 6226, 3728, 627 and 561 candidate proteolytic sites were detected in trypsin, V8 protease, chymotrypsin and CNBr digests respectively (using only N-terminal cleavages, as discussed in the previous section). 513 of these cleavage sites were found in two or more protease digests, including 28 that were found in three, and 3 that were present in all digests.

One of the 3 sites found in all digests is between A23-A24 in protein SO1164 (dacA-1). This site represents signal peptide cleavage site that was also predicted by SignalP and PrediSi [53, 98]. Another site detected in all digests is between A52-K53 in protein SO4509 (formate dehydrogenase, alpha subunit), which was previously detected in orthologous positions in two *Shewanella* species [51]. The third site is between T106-A107 in SO0417 which is annotated as putative pilin, and no prior knowledge is available for this protein. It is noteworthy that among the 513 sites present in two or more digests, 111 have A at P1 position, indicating the presence of many signal peptides, which are known to have a strong preference for A at P1 position in bacteria [53, 98].[2] Similarly, while one would expect false sites to be distributed uniformly across the lengths of the proteins, the sites with A at P1 position tend to appear in the first 40 positions (as expected for signal peptides). Given an average length of ≈ 300 residues for *Shewanella* proteins, we expect only 2 of the 513 sites to start at the second position of the proteins by chance. However, we find 55 cleavage sites at this position indicating the presence of many N-terminal methionine excisions (NME) [53]. Comparative analysis of multiple digests, therefore, is a promising approach for reliable identification of regulatory proteolytic sites.

One can observe that many cleavage sites detected by MS-Proteolysis belong to highly expressed proteins. For example, the translation elongation factor Tu (tufB) has so many identified peptides (230) that they result in appearance of 21 putative cleavage sites in tufB generated by MS-Proteolysis. However, while tufB is known to undergo proteolysis in bacteria [47], most of these sites are likely to represent artifacts rather than real proteolytic events. For example, various degradation variants of a highly ex-

[2] One would expect only ≈ 30 cleavage sites containing A at P1 by chance.

pressed protein may be detectable via MS/MS thus resulting in an artificial appearance of a cleavage site [121]. Since in vivo proteolytic events in such proteins are difficult to distinguish from artifacts, MS-Proteolysis generates an additional table that excludes all highly expressed proteins[3] and reports only the remaining peptides. Figure 6.5(a) shows the distribution of the starting positions of the detected cleavage sites and reveals pronounced peaks at the beginning of the protein (NME) and around position 25 (signal peptides). We further removed from consideration NME sites expected from NME specificity rules [53] and signal peptide cleavage sites predicted by SignalP and generated Figure 6.5(b) similar to Figure 6.5(a). Figure 6.5(b) still shows (a smaller) peak around position 25 indicating that SignalP failed to correctly predict some signal peptides. The peak becomes more pronounced, as shown in Figure 6.5(c), if we look at only the most upstream sites in proteins, indicating N-terminal proteolytic events like NME and signal peptide cleavage.[4] Therefore MS-Proteolysis is a useful tool for detecting the proteolytic events that software tools like SignalP miss. Since little is known about regulatory proteolysis in *Shewanella* apart from NME and signal peptides (CutDB database [63] does not report any proteolytic events in *Shewanella*), it remains to be verified which of these 175 putative cleavage sites represent in vivo proteolytic events. However, analysis of these sites reveals surprising biases that may warrant further studies. 33 out of these 175 sites have form A.* (19% of all sites) with surprisingly many A.A (11) and A.S (6) cleavages. While some of these sites may correspond to signal peptides missed by SignalP, others do not fit the profile of typical signal peptides and may reflect a still unknown proteolytic activity. Other surprisingly frequent cleavages are represented by Q.A and T.A (while these cleavages are expected to appear less than once by chance, they appear 6 and 7 times, respectively).

We also analyzed yeast(*S. cerevisiae*) proteome with MS-Proteolysis using trypsin digests. 7488 yeast peptides (identified with 0.1% FDR) yielded 11851 cleavage sites, of which 1047 were not explained by trypsin specificity and represented candidates for in vivo proteolytic sites. One of the tryptic cleavage sites is listed in CutDB database

[3] E.g., proteins that have over $threshold = 50$ identified peptides ($threshold$ can be set to a different value by the users depending on the levels of protein degradation in their samples)

[4] If the most upstream peptide identified in a protein (in a trypsin-digested sample, for example) starts at a non-tryptic position, it provides evidence for an N-terminal proteolytic event in the protein [53].

for *S. cerevisiae*, corresponding to the cleavage by protease Kexin between positions R40-Y41 in the protein Exg1p [6]. Having only a small overlap between annotated proteolytic sites in CutDB and sites revealed by MS/MS in yeast may be indicative of (i) limited coverage of proteolytic sites in CutDB, (ii) the fact that some proteolytic events are not represented in MS/MS sample since they appear only under specific conditions, and (iii) peptide detectability limitations in MS/MS analysis [88].

6.3 Discussion

Mass spectrometry is a reliable technology to determine the specificity of enzymes. We had previously demonstrated its application previously for trypsin [112], which was known to be very specific [100]. Here, we studied the specificity of V8 protease, chymotrypsin and CNBr), validated the known specificity rules, and found some interesting deviations from these known rules for the conditions used. Knowledge of these deviations is important for defining more accurate specificity rules for these proteases, and for analysis of regulatory proteolysis using mass spectrometry. Using comparative analysis of multiple proteases, we identified a set of putative in vivo proteolytic cleavage sites in *Shewanella*, which represent strong candidates for verification by future experiments. While some of these sites may represent various experimental and computational artifacts rather than than proteolytic cleavages, MS-Proteolysis represents the first step towards utilization of vast MS/MS datasets for studies of proteolysis.

Reliable peptide identifications are important for accurate determination of protease specificity from mass spectrometry. While we used ion-trap mass spectrometers to generate data for this study, we were able to keep the error rate extremely low by using doubly-confirmed cleavages or MS-GeneratingFunction. For future studies, using high precision instruments will be of additional help in detecting reliable cleavage sites.

Chapter 6 is, in part, a reprint of the submitted paper “Analyzing protease specificity and detecting in vivo proteolytic events using tandem mass-spectrometry. N. Gupta, K. K. Hixson, D. E. Culley, R. D. Smith and P. A. Pevzner”. The dissertation

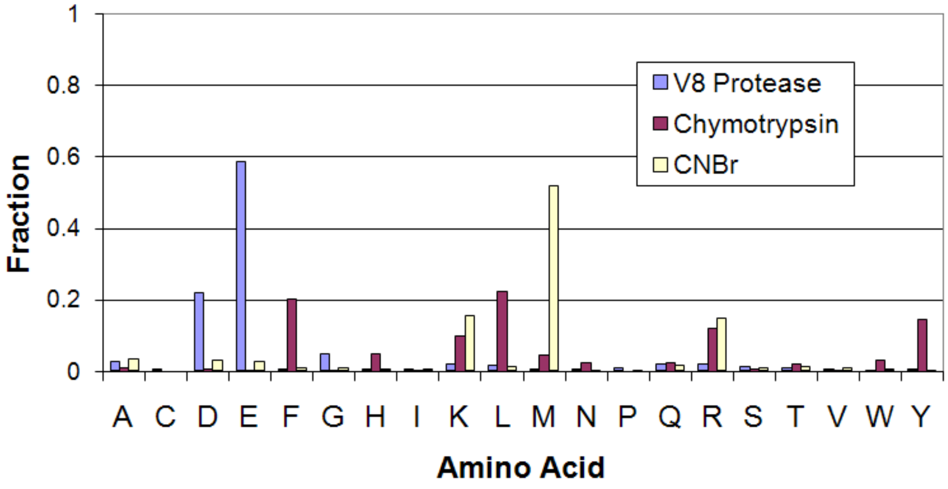
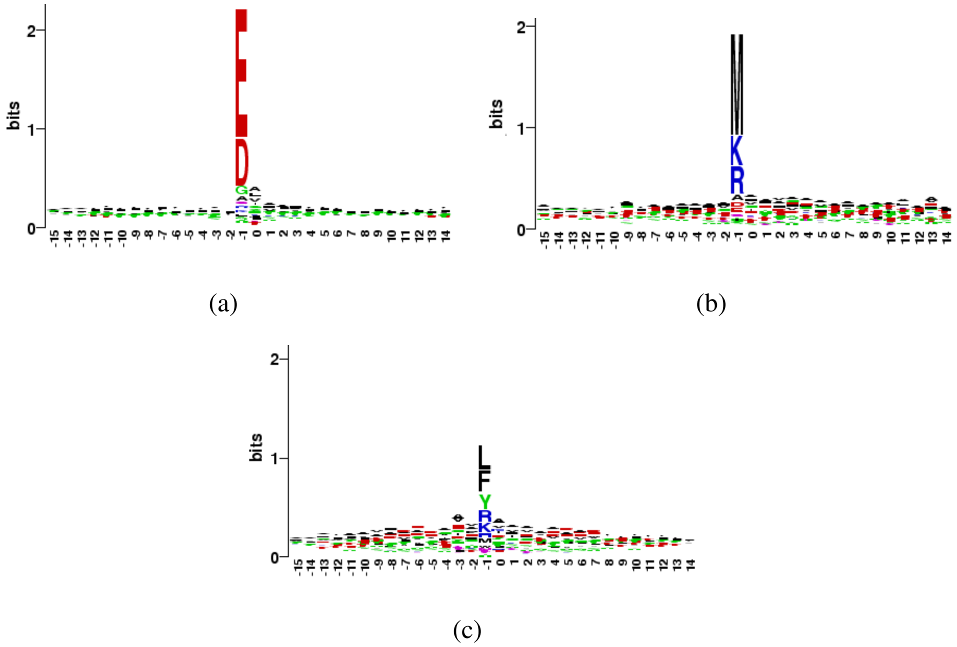


Figure 6.1: Fraction of different amino acids at P1 position in the doubly-confirmed cleavage sites, plotted for each of the three protease digests.

author was the primary investigator and author of this research.

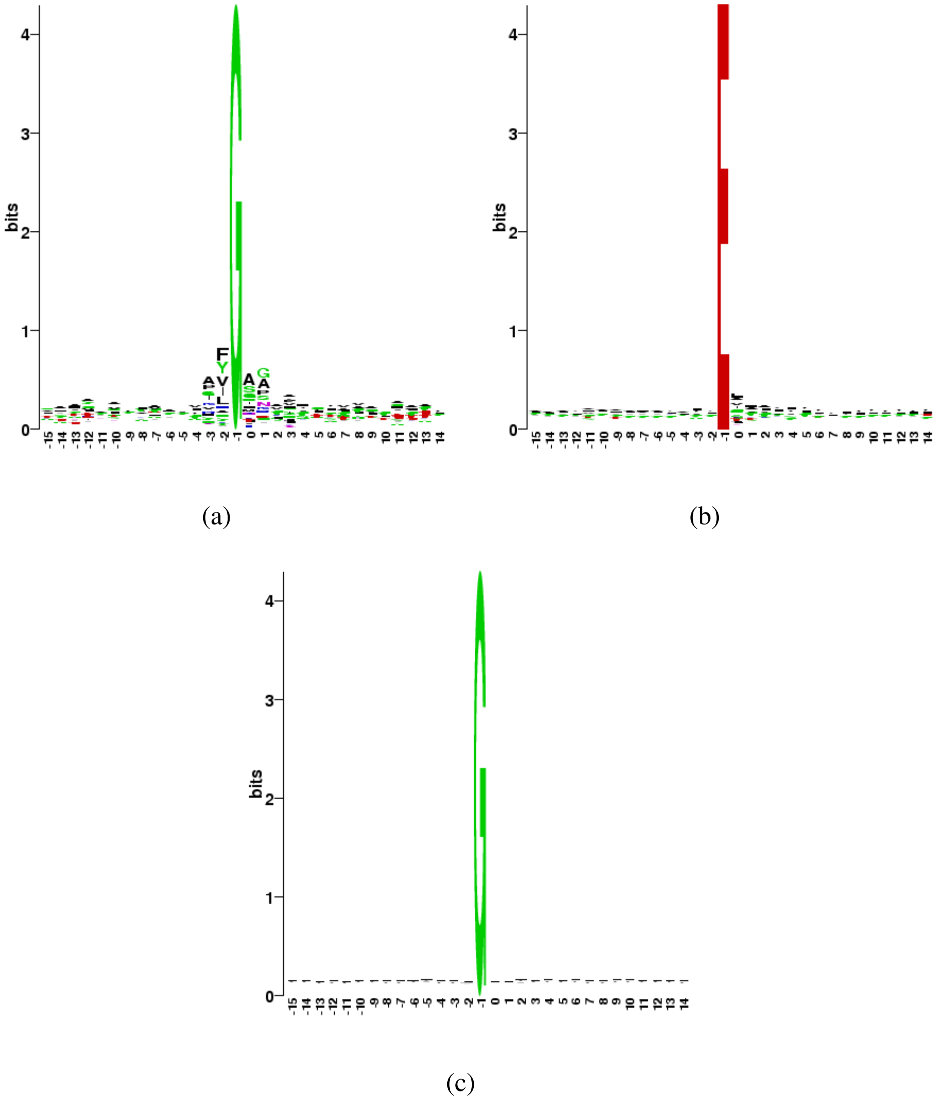


(a)

(b)

(c)

Figure 6.2: Sequence logo for the observed cleavage sites in (a) V8 protease (b) CNBr and (c) chymotrypsin. The P1 position is numbered -1 in the logos.

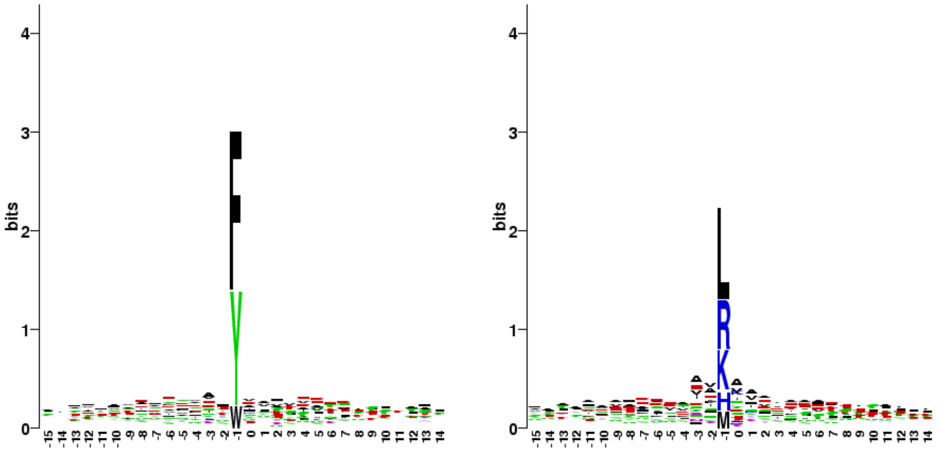


(a)

(b)

(c)

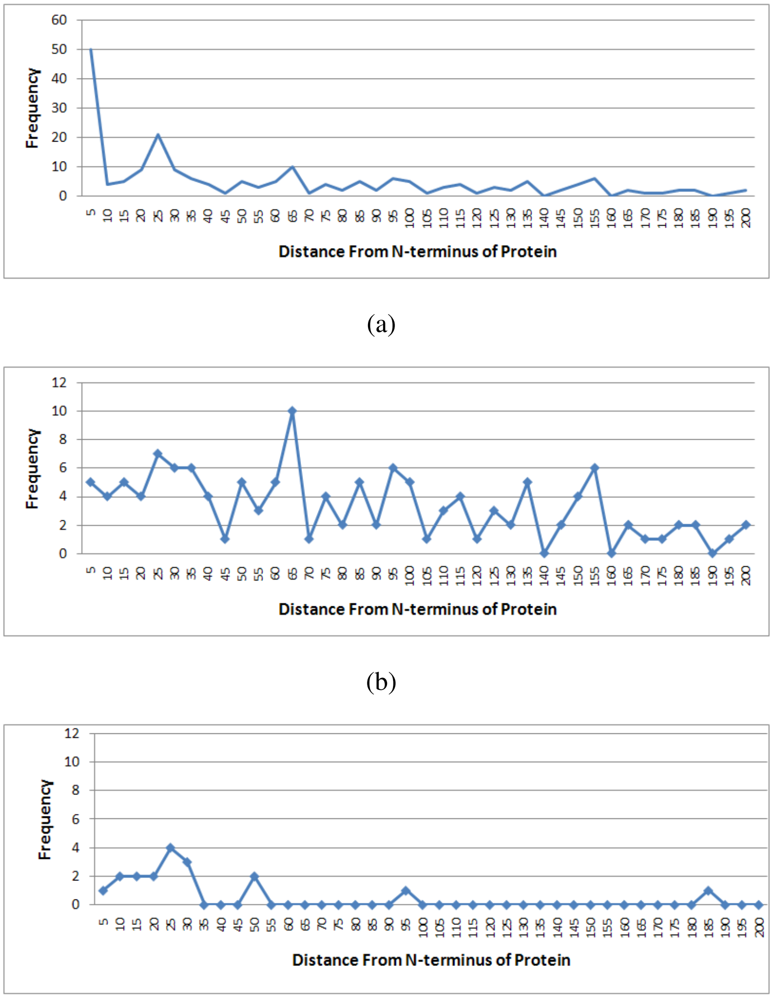
Figure 6.3: (a) Sequence logo for the observed cleavages in V8 protease digests at sites that have G at P1 position. (b) Similar logo for sites with E at P1 position. (c) Logo for all the 98,698 sites in *Shewanella* proteome that contain G (placed at -1 position in the logo).



(a)

(b)

Figure 6.4: (a) Sequence logo for the observed cleavages in chymotrypsin digests at sites that have F, W or Y at P1 position. (b) Similar logo for sites with H, K, L, M or R at P1 position.



(a)

(b)

(c)

Figure 6.5: **(a)** The histogram of positions in the corresponding protein sequence of the proteolytic sites. A bin size of 5 is used in the construction of histogram, and the plot is truncated at position 200 for brevity. **(b)** Similar plot as in (a), after removing the sites at second positions of proteins (NME) and those predicted as signal peptide cleavage sites by SignalP. **(c)** Similar plot as in (b), but keeping only the most upstream peptide detected in a protein (to infer N-terminal proteolytic events like NME or signal peptide cleavage [53]).

Table 6.1: Frequency of different amino acids at P1 position in the double-confirmed cleavage sites observed for the three protease digests. The last column indicates the background frequency (count) of each amino acid in the entire *Shewanella* proteome. The amino acids defining the commonly accepted specificity for V8, chymotrypsin, and CNBr are shown in bold. The amino acids that do not contribute to known specificity rules but have surprisingly large counts in P1 positions of the cleavage sites are shown in italics.

Amino Acid	V8 protease	Chymotrypsin	CNBr	Background
A	215	22	30	136659
C	20	0	0	16251
D	**1894**	10	26	77030
E	**5044**	4	24	83281
F	47	**431**	6	57812
G	*393*	4	8	98698
H	23	*101*	3	34257
I	29	6	3	88040
K	146	*209*	*135*	75790
L	135	**476**	10	159360
M	39	*96*	**448**	37920
N	45	50	2	60337
P	53	0	1	59308
Q	164	48	13	71641
R	158	*255*	*127*	68627
S	90	10	8	95067
T	72	42	11	78882
V	32	6	6	98094
W	3	**65**	3	18888
Y	33	**311**	2	44781
Total	8635	2146	866	1460723

Chapter 7

Novel proteogenomic applications

7.1 Introduction

This chapter presents some specific applications of proteogenomic approaches that have been developed in previous chapters. In particular, we demonstrate how the approaches developed in this work for analyzing proteolytic sites may be adapted for analysis of novel neuropeptides, for discovery of mutations and rare modifications using high-resolution MS/MS data. The last section discusses how mass spectrometry based evidence may be useful in improving operon predictions which has proved to be a difficult problem in the field of genomics.

7.2 Identification of neuropeptides

Neuropeptides are important regulators of several neurological and neuroendocrine physiological processes including pain, anxiety, behavior and metabolism. However, our knowledge of actual neuropeptide sequences and the mechanisms of their production and regulation is limited. Neuropeptides are produced by proteolytic cleavage from proproteins targeted for secretion and regulation of neurotransmitter functions.

To understand the proteolytic mechanisms underlying the production of neuropeptides, we ask two important questions: (1) What cleavage sites participate in in-vivo processing and what protease specificities would account for these cleavages? (2)

What peptide products are produced from these in-vivo cleavage events? In this study, we explore the use of mass spectrometry as a high-throughput approach to analyze the neuropeptidome in human and bovine chromaffin granules.

Mass spectrometry is well suited for unrestricted analysis of neuropeptides in neurological and neuroendocrine cells and tissues [80, 127]. This technique provides direct evidence for peptide sequence, N- and C-terminal extensions and site-specific information on post-translational modifications without a priori knowledge of the target peptide. Mass spectrometry also has the advantage of being a high-throughput technology capable of separation and analysis of highly complex biological samples. In contrast, radioimmunoassay (RIA) is an antibody-based technique that requires precise knowledge of the target sequence. RIA does not identify the molecular form of the target peptide, rather it detects the presence of related peptide sequences that bind to the antibody. Therefore, N- or C-terminal extensions or modifications may not be observed by RIA techniques. Furthermore, amino acid sequences between active neuropeptides and intermediates are not observed by current RIA methods. This is important because the best evidence for specific proteolysis of prohormones would be most indelibly etched into these intervening peptide sequences.

7.2.1 Methods

Low molecular weight peptides were filtered from chromaffin granules from a human adrenal pheochromocytoma tumor and normal bovine adrenal medulla. The human samples were analyzed under three conditions: without digestion, with trypsin digestion and with V8 protease digestion. The samples were subjected to LC-MS/MS analysis after solid phase extraction. Data sets were generated on an ion-trap instrument as well as a high resolution QTOF instrument. The peptide sequences were identified with InsPecT against the human and bovine IPI databases. Since only a very small fraction of proteins in these databases are expected to be present in the granules, we employed a two-stage database search, where the IPI databases were initially searched to prepare smaller databases containing the likely proteins, which were rigourously searched again with a 1% false discovery rate threshold measured through decoy databases.

7.2.2 Results

Peptides identified in the undigested samples provided direct evidence for in-vivo proteolytic cleavages (after filtering out in-vitro degradation products of peptides). A total of 181 peptides (from 16 proteins) were identifed in human and 138 peptides (from 22 proteins) were identified in the bovine samples. Neuropeptides are commonly known to be flanked by dibasic amino acid motifs, as confirmed in Figure 7.1. We also found a number of peptides that deviated from this rule, and these showed an abundance of acidic amino acids (D and E), as shown in Figure 7.2. While the acidic amino acids are spread out around the cleavage position in human samples, the bovine samples show a remarkable preference for D at P1 position, alluding to a new cleavage mechanism for these peptides. Significant evidence for other proteolytic cleavage sites was further observed, with at least 5 representative sequences of R/K-Xn-R/K where n = 1 or 2 observed in proteins CgA, CgB and NPY alone. Amino-tripeptidyl cleavages were also detected in this tissue.

In this study, we are not only interested in identifying proteins present in the samples, but also identifying the individual peptides to better understand the processing of these proteins. Each enzymatic condition, in the digestion step of sample preparation, allows identification of a certain set of peptides. The physiochemical properties of peptide sequences make some of these peptides less detectable than others by mass spectrometry [88]. To capture as many peptides as possible, we experimented with using multiple enzymes for digestion of the human samples. In addition to running the low molecular weight fractions without digestion, the samples were also digested with trypsin and V8 Protease. Figure 7.3 shows that majority of the peptides identified in any enzyme are unique to that condition (relatively small overlaps). Thus using the three enzymes in conjunction significantly increases the overall number of peptide identifications (from 181 in no-protease samples to 330 in total). The number of proteins identified also increases by using multiple enzymatic conditions from 16 to 23.

Using multiple enzymes also results in better coverage of the proteins, as exem-

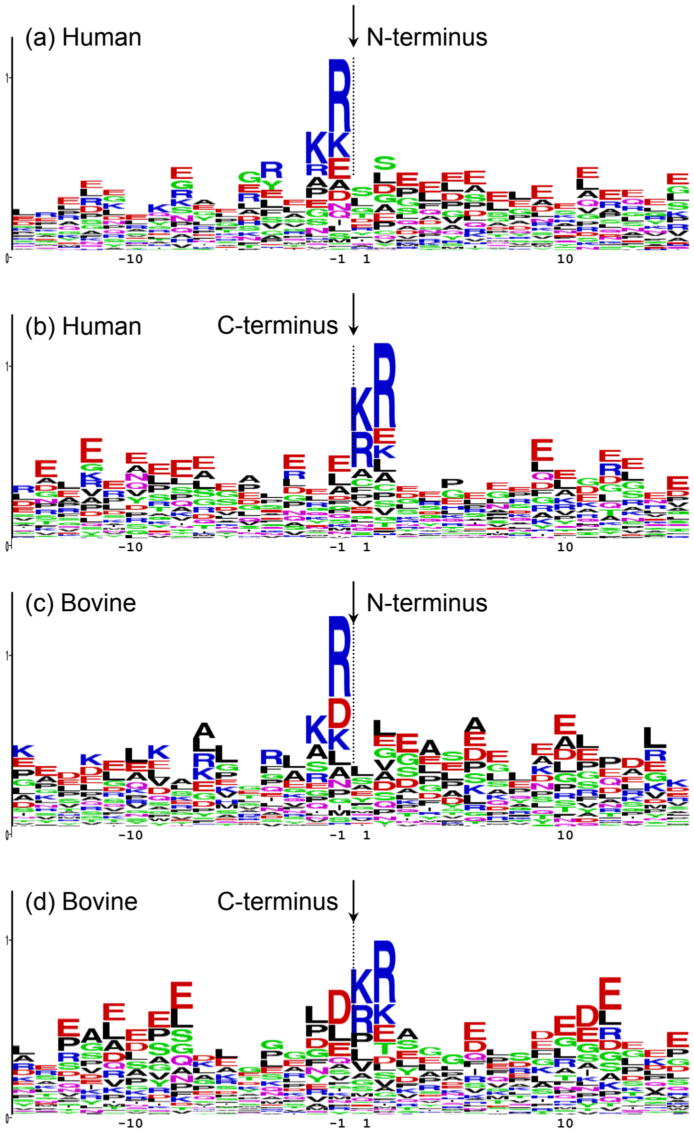


Figure 7.1: Sequence logos for the N- and C- termini of the observed neuropeptides in the undigested samples. Fifteen amino acids are shown on either side of the cleavage site (between positions -1 and 1 in the above figures).

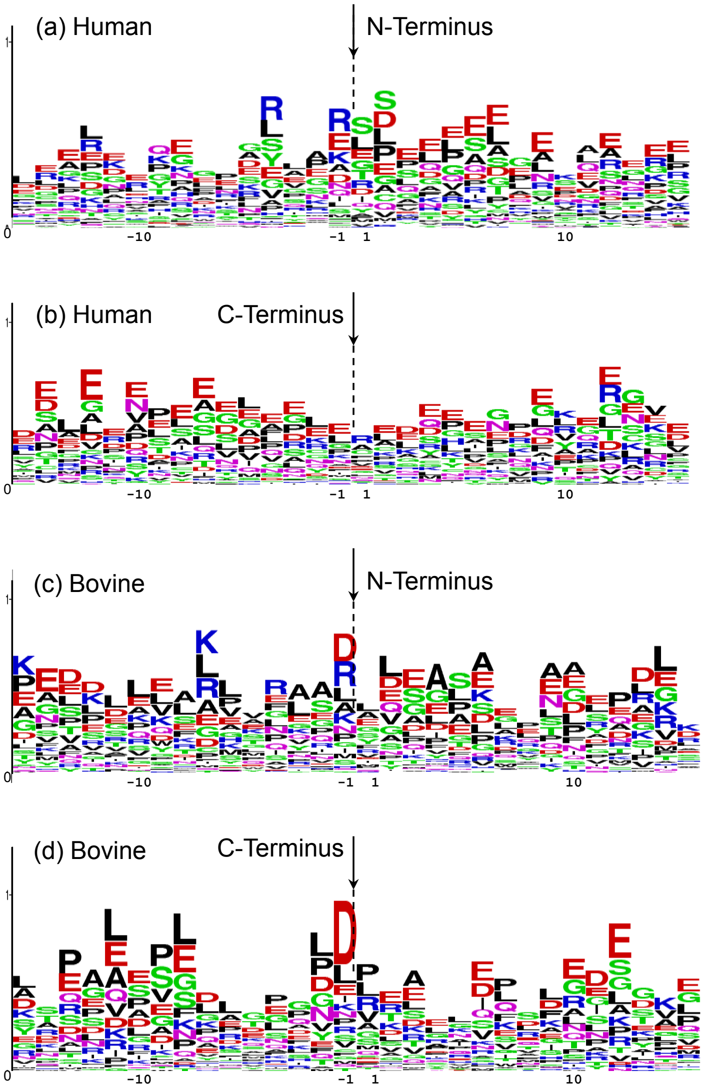


Figure 7.2: Sequence logos for the N- and C- termini of the observed neuropeptides in the undigested samples, after removing the sites explained with dibasic amino acids.

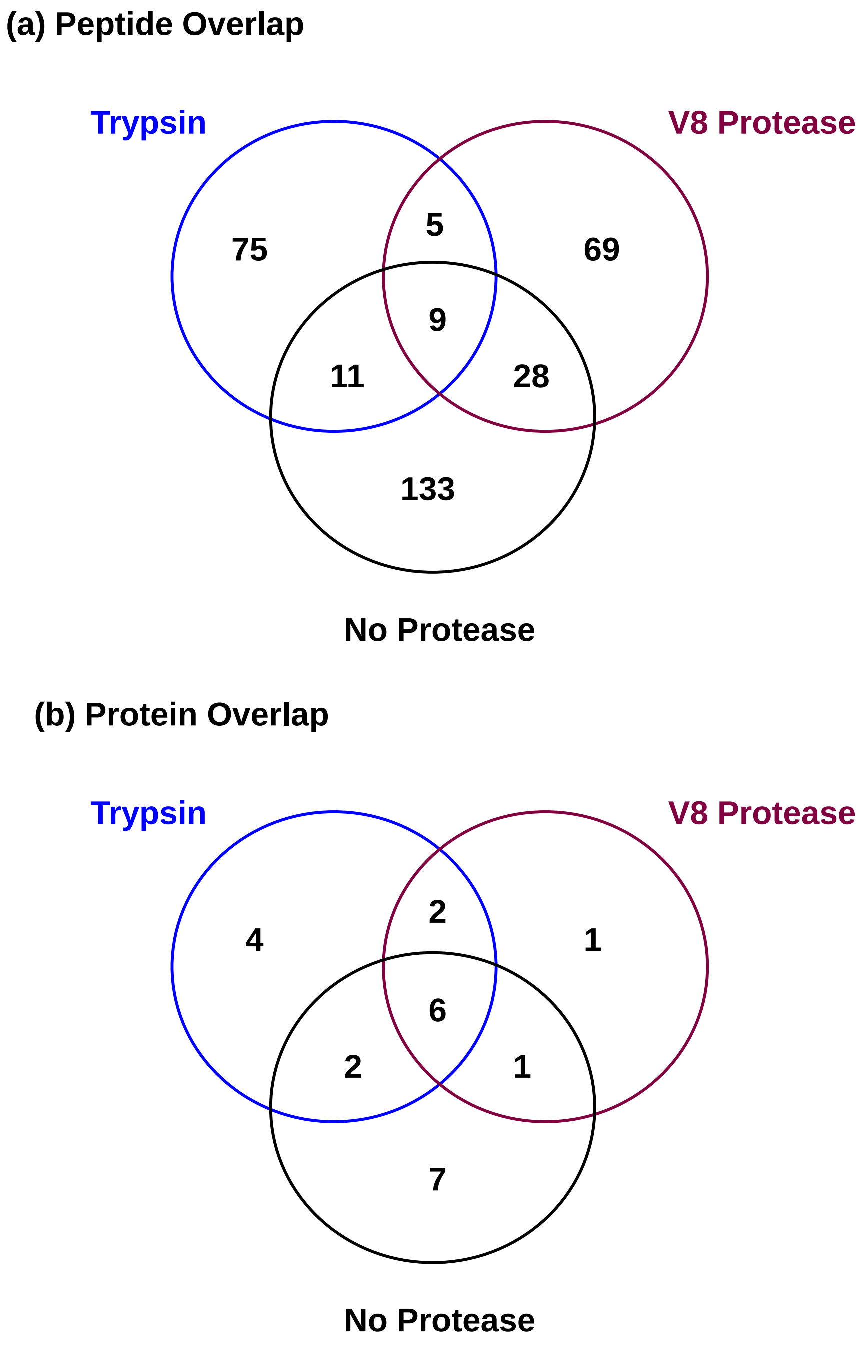


Figure 7.3: Overlap in the number of peptide and protein identifications in three human samples. A total of 330 unique peptides and 23 proteins are identified in the three experimental samples.

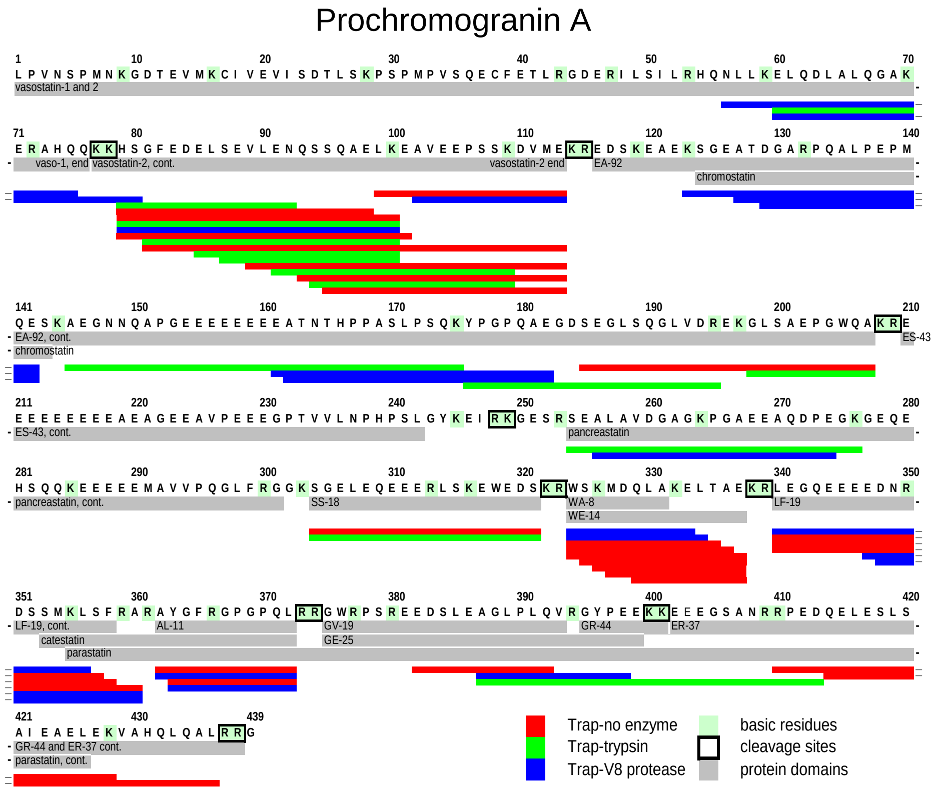


Figure 7.4: Peptides identified by InsPecT in prochromogranin A at 1% FDR level, from ion-trap data set, analyzed in three enzyme-conditions (no-enzyme, trypsin digests and V8 protease digests) shown in distinct colors. The basic amino acids in the protein sequence are colored in light green, previously known cleavage sites are shown in blocks, and previously known neuropeptides are shown in yellow. Hyphens at the end of some lines indicate peptides that got split between two lines in making the figure.

plified by Prochromogranin A (Figure 7.4). 36.4% of the protein sequence is covered by endogenous peptides (without enzyme digestions) identified by InsPecT. However, when peptides identified from trypsin and V8 digestions are included, the coverage increases to 64.5%. Two neuropeptides - Chromostatin and Pancreastatin - were detected only with trypsin and V8 digestions. Similar increase in coverage is seen in all proteins that we analyzed. We therefore advocate the use of multiple enzymes routinely in peptidomic studies.

Conspicuously absent in the proenkephalin data obtained from the ion-trap instrument were enkephalins. These are small neuropeptides comprised of the sequences YGGFL (leu-enkephalin) and YGGFM (met-enkephalin) known to be produced from this protein. These neuropeptides are 5 amino acids in length and provide few fragment ions when subjected to collision-induced dissociation. While it was anticipated that low-resolution fragmentation data from the ion-trap instrument would be insufficient to identify enkephalins effectively in complex database search analysis, high-resolution data could surmount this issue. As evident in Figure 7.5, the high-resolution data from the Agilent QTOF instrument proved sufficient to unambiguously identify met-enkephalin directly from database search analysis. At a lower stringency level of 5% FDR (instead of usual 1%), leu-enkephalin was also detected.

7.2.3 Discussion

We found a surprising number of peptide cleavages after D, particularly in bovine samples. A few other striking observations about neuropeptides can be made from this data: (i) We find that intervening sequences between known neuropeptides are, counter-intuitively, seen with high peptide coverage. (ii) Evidence is seen for novel putative neuropeptides with the canonical dibasic motif.

We also find that the inclusion of trypsin and V8 protease digestions significantly increases the number of unique peptide identifications (different lengths and endpoints), providing better insights into the processing of proproteins and the production of neuropeptides. Our results indicate that the use of multiple enzymes for protein digestion is a powerful, albeit under-utilized, approach for improving discovery rate in mass spec-

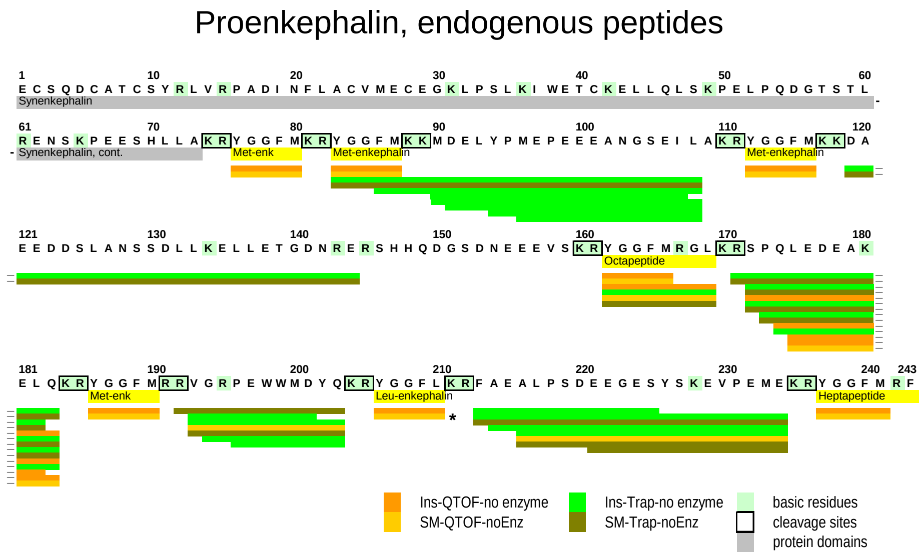

Proenkephalin, endogenous peptides
1 10 20 30 40 50 60
E C S Q D C A T C S Y R L V R P A D I N F L A C V M E C E G K L P S L K I W E T C K E L L Q L S K P E L P Q D G T S T L -
Synenkephalin
61 70 80 90 100 110 120
R E N S K P E E S H L L A K R Y G G F M K R Y G G F M K K M D E L Y P M E P E E E A N G S E I L A K R Y G G F M K K D A
- Synenkephalin, cont.
Met-enk
Met-enkephalin
Met-enkephalin
121 130 140 150 160 170 180
E E D D S L A N S S D L L K E L L E T G D N R E R S H H Q D G S D N E E E V S K R Y G G F M R G L K R S P Q L E D E A K
Octapeptide
181 190 200 210 220 230 240 243
E L Q K R Y G G F M R R V G R P E W W M D Y Q K R Y G G F L K R F A E A L P S D E G E S Y S K E V P E M E K R Y G G F M R F
Met-enk
Leu-enkephalin
*
Heptapeptide
Ins-QTOF-no enzyme
SM-QTOF-noEnz
Ins-Trap-no enzyme
SM-Trap-noEnz
basic residues
cleavage sites
protein domains

Figure 7.5: Schematic display of endogenous peptides identified in preproenkephalin. Peptides identified through ion-trap data as well as QTOF data, with InsPecT (Ins) and SpectrumMill (SM) at 1% FDR are shown in distinct colors below the protein sequence (in black). The basic amino acids in the protein sequence are colored in light green, previously known cleavage sites are shown in blocks, and previously known neuropeptides are shown in yellow. Hyphens at the end of some lines indicate peptides that got split between two lines in making the figure. *Leu-enkephalin was detected only at 5% FDR.

trometry especially when the sequences of identified peptides are important, as opposed to just the identity of proteins.

7.3 Using high-resolution MS/MS data for detecting mutations and rare modifications

What do post-translational modifications (PTMs), chemical adducts, amino acid mutations and sequencing errors have in common? They all can modify the mass of a peptide analyzed by mass spectrometry, compared to the mass of its sequence in the database. Indeed, spectral data sets obtained in almost every mass spectrometry experiment contain such wide variety of information. Yet, we have been limited by the existing computational approaches in exploiting this gold-mine. Analysis of modified peptides has traditionally been limited to specific modifications of interest (such as phosphorylation).

Recent development of tools, such as MS-Alignment [136] and ModifiComb [117], that enable unrestricted (*blind*) search for modified peptides alleviates this problem to some extent. However, these tools often reveal not only a large number of modifications, but a large variety of them, leading to a new challenge in interpreting the results. To avoid calling computational artifacts as real modifications, the studies describing MS-Alignment or ModifiComb used a “strength-in-numbers” approach to select reliable modifications [136, 117]. While this approach allows one to identify common PTMs or chemical adducts (such as oxidations or pyroglutamate), it has limitations in identifying mutations and rare in vivo PTMs.

Assuming each amino acid can mutate into one of the other 19 amino acids, there are 380 possible types of mutations in nature. However, the actual number of mutations (or sequencing errors) present in any particular proteome sample analyzed by mass spectrometry is usually smaller. Therefore, any particular type of mutation detected in a blind search of modifications will typically be seen with a very low frequency, making it difficult to distinguish it from computational artifacts. Moreover, the mass shift corresponding to a mutation may coincide with the mass shift of a rare modification, making it difficult to decide whether a modified peptide actually represents a mutation or a PTM. In this study, we analyze high-accuracy spectra from *Shewanella oneidensis* MR-1 and *E. coli* and develop a new computational approach for reliable identification of mutations.

In the seoncd part of this study, we use comparative analysis of different bacterial species to identify rare PTMs. Important PTMs like diphthamide (observed on histidine in translation elongation factor 2) might be seen only once in the entire proteome, also making them difficult to distinguish from noise. In Gupta et al. 2008 [51], we described a comparative analysis of three *Shewanella* species to identify conserved rare modifications. However, the analysis had the limitation of using only very closely related species (all *Shewanella*) and low-accuracy ion-trap data. To avoid making false predictions, the analysis was limited to modifications that were seen at exactly the same position on two or more orthologs. Here, we demonstrate that using high-accuracy data and comparing with *E. coli*, a more distant relative of *Shewanella*, enables us to bypass this restriction and increase the number of confident predictions.

7.3.1 Methods

The *Shewanella* dataset comprised 2 million high-accuracy spectra generated on an FT-ICR instrument at Pacific Northwest National Laboratory, previously described in Gupta et al, 2007 [53]. The spectra were derived from *Shewanella oneidensis* MR-1 cells grown under different various experimental conditions. *E. coli* dataset comprised 0.7 million high-acccuracy spectra generated in the same laboratory. The accuracy of the parent mass for these spectra is expected to be 5ppm or less, i.e., usually less than 0.01 Dalton.

MS-Alignment [136] was used to identify peptides with unknown mass shifts in the range of -200 to +250 Daltons. At most one modification was allowed on each peptide. The searches, for each organism, were carried out against a database containing the protein sequences of that organism and a decoy database of the same size containing shuffled protein sequences.

MS-GeneratingFunction allows accurate computation of the false positive rate (FPR) of individual peptide-spectrum matches, and has been shown to result in a better sensitivity-specificity trade-off compared to traditional database search scoring functions for unmodified peptides [75]. Here, we extended MS-GeneratingFunction to handle peptides with modifications by treating the modified residue as a new type of amino acid, and computing the *spectral probabilities* in the 21 amino acid alphabet which could be different for different peptides. While this is a gross simplification, it is acceptable for this study where the goal of using MS-GeneratingFunction is only to filter out low quality identifications that spuriously got high scores in database search, and we . At a spectral probability threshold of 6×10^{-10}, 30009 peptides from MS-Alignment results were selected in *Shewanella* and 6499 peptides were selected in *E. coli* at 0.005 false discovery rate (FDR), estimated using the decoy database.

Unmodified peptides reported by MS-Alignment were removed. MS-Alignment reports modifications as integer mass shifts. To take advantage of our high-accuracy data, we derived accurate masses of the shifts by subtracting the sum of exact masses of all amino acids in the peptide from the parent mass of the peptide. Positive mass shifts that could be explained by a missing amino acid at one of the termini, or negative mass shifts that could be explained as an extra amino acid at one of the termini, were not included in the analysis.

We used ortholog clusters, obtained from the eggNOG database [66], to infer ortholog-mappings between the two bacteria. The eggNOG database contains orthologous groups constructed from SmithWaterman alignments through identification of reciprocal best matches and triangular linkage clustering. We selected all available COG mappings for *Shewanella* and *E. coli* proteins from this database. Proteins belonging to the same clusters were treated as orthologs.

7.3.2 Mutations

All modified peptides identified by MS-Alignment can be summarized in the form of a *frequency matrix* [136]. Note that we use the term *modification* to represent any observed mass shift on a peptide, irrespective of whether it represents an in vivo PTM, a chemical artifact added during mass spectrometry, a mutation in the protein sequence or a sequencing error. Each row in the matrix represents a possible mass shift (we use a resolution of 0.01 Daltons here for analysis of mutations), in the range of -130 to +130 Daltons, enough to allow the mass difference between Gly and Trp, the lightest and the heaviest amino acids. Each of the 20 columns in the matrix represents an amino acid on which the mass shift is observed. The frequency matrix, therefore, contains $260 \times 100 \times 20 = 520000$ cells. The matrix is sparse, with heavy clusters usually seen around some well-known PTMs.

Each of the 20 amino acids can possibly mutate into any of the other 19 amino acids. Note that L and I have the same mass, therefore effective number of possible mass shifts is 18 for any amino acid, leading to 360 possible mass shifts that correspond to mutations. Since we assume an accuracy of 0.01 Daltons in our measurement of mass shifts, a mutation an amino acid may fall into one of the two cells on adjacent rows for the corresponding column. Therefore, 720 cells in the frequency matrix are designated as *putative mutation cells*.

Table 7.1: Frequency of different types of cells in the frequency matrix (*cell* here stands for an element of a matrix, not to be confused with the biological meaning of cell). The second column indicates the total number of cells of different types, and the third column indicates their total frequency. Since a row in the frequency matrix represents a 0.01 Dalton window, each $Mutation + 5H$ cell is the cell 504 rows below the corresponding putative mutation cell ($Mutation - 5H$ cells are defined similarly in the other direction).

Cell type	Number of cells	Total frequency
All	520000	2053
Putative Mutation	720	103
$Mutation + 5H$	720	4
$Mutation - 5H$	720	5

A total of 103 peptides in *Shewanella* map to the 720 putative mutation cells. To determine if this number is statistically significant, we compare it against the total frequency of arbitrarily chosen 720 cells in the frequency matrix (Table 7.1). Since the matrix does not have a uniform distribution (smaller mass shifts are more prevalent than large mass shifts), randomly selecting 720 cells in the matrix might not provide an accurate control. Instead, we select the 720 cells for comparison by finding the cells that correspond to mass shift values of rougly 5 Daltons higher than actual putative mutation cells. Another concern might be that an artificially selected mass shift may lead to a biochemically infeasible composition and will therefore not provide an accurate control. To avoid this issue, we use cells shifted by 5.04 Daltons (five times the mass of Hydrogen atom) relative to the putative mutation cells. The control cells thus chosen are labeled as $Mutation + 5H$ cells. These cells exhibit a combined frequency of 4 in *Shewanella*. Similarly, $Mutation - 5H$ cells, selected by finding cells with mass shifts 5.04 Daltons smaller than the putative mutation cells, show a frequency of 5. Thus, we could identify 103 candidate mutations with less than 5% error rate. Similar results are seen in the *E. coli* data set.A total of 112 mutations are identified, with an error rate estimated to be between 2% and 20% from the $Mutation + 5H$ and $Mutation - 5H$ cells.

We further suggest a filtering approach to select a subset of these putative mutations with extremely low error rate (albeit with lower sensitivity). This filtering makes use of the high resolution of our MS/MS data, and the observation that many mutations are found in both directions, i.e., if a mutation from amino acid X to Y is seen with mass shift δ, there might be also be a mutation from Y to X with mass shift $-\delta$ on a different peptide. This occurrence of *pairs* of mutations, with exactly the same magnitude of mass shifts but in opposite directions, is not expected for computational artificats or peptides representing PTMs, especially if one uses high resolution windows (0.01 Daltons) to specify mass shifts, as done here. As shown in Table 7.2, there are 27 cells in the frequency matrix corresponding to the paired mutations in *Shewanella*, with a total frequency of 37. In contrast, the cells that are 5.04 Daltons above and below these cells do not contain any peptides, indicating that the 37 candidate mutations are almost error-free. In contrast, no paired mutations are detected in *E. coli*, indicating that while the paired-mutation requirement can be useful where very high specificity is desired, it is not practical in all circumstances because of low sensitivity.

Table 7.2: Modified version of Table 7.1, showing only the *paired* mutation cells for *Shewanella*, i.e., including only mutations that are seen in both directions. Equal number of control cells, $Mutation + 5H$ and $Mutation - 5H$, are chosen for these cells.

Cell type	Number of cells	Total frequency
All	520000	6695
Paired Mutation	27	37
$Mutation + 5H$	27	0
$Mutation - 5H$	27	0

In the preceding discussion, we put no restrictions on which mutations are allowed. However, in nature, mutations may be more likely between selected pairs of amino acids (such as changing from one hydrophobic amino acid into another) compared to others. PAM substitution matrices reflect the rates of mutation between pairs of amino acids at different evoluationary time scales. The average distance between all possible 380 pairs of amino acids in PAM2 matrix (corresponding to the shortest possible evolutionary distance, appropriate for the mutations observed here) is -16.0. In comparison, the average distance between the pairs of amino acids in the 103 mutations observed in *Shewanella* is -11.8, a lower distance indicating that more mutations were indeed expected between these pairs of amino acids. Similarly, the average distance for the 112 mutations detected in *E. coli* is -11.4, lower than the average distance in the matrix. These results provide additional support to the detected mutations by showing that they agree with expected rate of substitution between amino acids.

For three of the mutations identified in *Shewanella*, we see an ortholog in another *Shewanella* species containing the substituted amino acid at the same position. For instance, in protein fdhB-2 *Shewanella oneidensis* MR-1, E78 is seen with a mass shift of -14.02, corresponding to a mutation to D. The orthologous proteins in *Shewanella putrefaciens* CN-32 and in *Shewanella sp.* W3181 have a D at the same position.

7.3.3 Rare PTMs

1118 and 606 cells are populated in the PTM frequency matrices in *Shewanella* and *E. coli* respectively, out of 520000 cells in each. If the same number of cells were randomly picked from the matrices, the expected number of cells that would be common between them is less than 2 [1]. In reality, however, we find that 144 of the populated cells in the two matrices are common, indicating that these conserved cells represent reliable modifications as opposed to computational artifacts.

We find that many of these 144 common modifications are found on orthologous proteins in *Shewanella* and *E. coli*. Assuming that we have $K \approx 3000$ orthologous clusters (see Methods) between the two organisms, the expected number of ortholog pairs in each cell of the frequency matrix is given by $n_1 \times n_2/K$, where $n1$ and $n2$ represent the number of peptides carrying the modification in the two organisms respectively. Therefore, the total expected number of ortholog pairs can be computed as the

[1] This estimate (1118 $\times$ 606/520000 $\approx$ 1.3) assumes a uniform distribution of populated cells in the matrix for simplicity of calculation. However, the expected number would still smaller than the observed 144 even if one takes the uneven distribution into accounts.

sum of this fraction over the 144 cells, approximately 13. The observed number of orthologous pairs is 130, significantly larger than the number expected by chance, further providing evidence in favor of the reliability of these modifications. Table 7.3 provides a summary of the modifications that were conserved in orthologous proteins.While some of the promiment modifications in this list are well known and expected (such as oxidation of Methionine, or pyroglutamte), a few are surprising and may represent rare PTMs. For example, -18.01 modification in translation elongation factor Tu seen on the first leucine residue in peptides E**L**LSEYDFPGDDLPVIQSGSALK in *Shewanella* and E**L**LSQYDFPGDDTPIVR in *E. coli* suggests that this modification is evolutionarily conserved.

Table 7.3: Number of ortholog pairs on which different PTMs were found to be conserved between *Shewanella* and *E. coli*.

Amino acid	Mass shift	Ortholog Pairs
E	15.99	10
E	31.99	1
H	-17.03	1
I	-17.03	4
K	-17.03	1
L	-18.01	2
L	-17.03	1
L	-17.02	2
M	15.99	88
M	31.99	6
P	15.99	1
Q	-17.03	4
Q	15.99	2
R	-17.03	1
V	-17.03	1
V	-17.02	1
V	15.99	2
W	15.99	1
W	31.99	1

7.4 Improving operon predictions using mass spectrometry

Operon predictions reveal co-regulated genes, a crucial step towards understanding the regulatory pathways of the cell. Most methods for predicting operons rely solely on sequence information such as the distance between consecutive genes or functional annotations [18, 149, 115, 14, 108]. Another method previously applied for operon predictions is the use of conserved gene clusters [18,38], assuming that the genes belonging to the same operon will remain clustered in related species. Operon prediction, however, has proved to be a difficult problem. In their recent review of various operon prediction algorithms, Brouwer et al., 2008 concluded that there is much room for improvement, and suggested that experimental evidence such as from transcriptomics can be helpful in improving the accuracy of predictions [18]. Previous efforts to detect operons using experimental evidence have been carried out with DNA microarray experiments [18]. This involves examining the expression profiles of genes under different growth situations to identify adjacent genes that are co-regulated.

The sequence-based methods of operon prediction are only as good as the underlying models. Since these are usually trained on a few specific organisms, their performance on other organisms is not always satisfactory. On the other hand, using experimental evidence, such as from microarrays, for determining operons gives more accurate results when applied to any organism, but these experiments are rather time consuming and expensive. Therefore, accessibility to other high-throughput experimental methods, if demonstrated to be useful for operon predictions, can be valuable as a complementary approach to microarrays.

Advances in tandem mass spectrometry (MS/MS) now allow for rapid detection of peptides from whole proteomes [1]. Peptide identifications from MS/MS database searches provide clues for protein expressions, and indirectly, for gene expressions. Since an operon produces a polycistronic mRNA, its protein products are expected to be expressed in synchrony. We compared the operons in *Pyrococcus furiosus* predicted by the sequence-based and microarray-based approaches [133] with experimental data from tandem mass spectrometry, and checked if the protein expressions determined from mass spectrometry are consistent with the putative operons. Though the proteome gives only indirect evidence towards operon prediction as compared to mRNA analysis, we show that it is possible to derive valuable evidence for confirming and correcting the operon annotations using proteomic data.

7.4.1 Results

We present a software tool, MS-Operon, which combines proteomic information with sequence-based operon predictions and helps in correcting operon annotations. Analysis in MS-Operon is carried out through queries – searches that find operons with different type of expression patterns. MS-Operon includes queries with three basic patterns of protein expression within imported operon sets. After an operon set is loaded into MS-Operon, searches can be made into the dataset to select and analyze operons with specific types of protein expression patterns. Users can also define the maximum and minimum operon length for the search. We denote each protein by a 0 if it is not supported by any peptide and by 1 if at least one supporting peptide is found. Thus an operon can be written as a string of 1s and 0s. MS-Operon conducts three basic searches: All Expressed (all 1s), None Expressed (all 0s), and Split (a run of 1s followed by a run of 0s, or the converse). A fourth search option allows visualization of all operons overlaid with protein expression information.

Whole proteome MS/MS spectra were generated from the archaeon *Pyrococcus furiosus* (spectra have been previously reported in [134]. InsPecT [126] identified 16252 peptides in the dataset at a false discovery rate of 0.01 measured using the standard target-decoy approach. We compared the mass spectrometry data with two different operon sets for *Pyrococcus furiosus* published by Tran et al., 2007 [133]. The first set was derived from microarray data, and the second set was from a neural network (NN) predictor that combined results from three different previously published sequence-based operon prediction algorithms [133].

MS-Operon confirmed many predicted operons, as well as provided evidence for correcting other operons in Pyrococcus furiosus. Of the 50 operons in the microarray-predicted dataset with length 5 or more, 13 (26%) are seen in the Split search and are possible candidates to be split into two operons or shortening of operons. At the

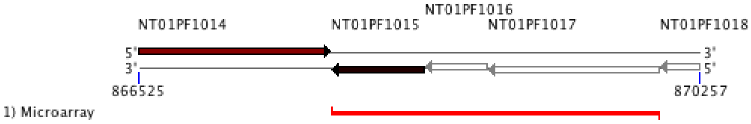


Figure 7.6: Graphical illustration of a putative operon in an MS-Operon report. The pointed arrows, drawn on the two complementary strands of the chromosome, represent the genes in the operon and their directions (protein names are indicated above). The coordinates of the region of the chromosome are marked by the two numbers at the ends. The position of the operon is indicated by the red line (user assigned name for the operon set is indicated on the left), and one flanking gene on either side of the operon is shown on the chromosome for reference. Arrows are white when no peptide is found for the corresponding protein, and black through red for an increasing number of supporting peptides. In the particular case shown here, an expressed gene (NT01PF1015) within the operon followed by two contiguous non-expressed genes indicates a possible error in the predicted operon (see Results for details).

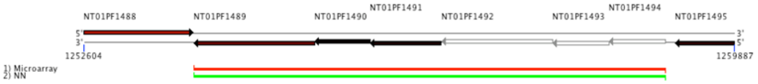


Figure 7.7: Operon NT01PF1489 - NT01PF1494 in *Pyrococcus furiosus*. Two operon sets, viz. microarray-determined operon set (red line) and the neural network (NN)-determined operon set (green line), were compared with mass spectrometry data. The first three genes in the operon show high expression (evident from multiple peptides), while the last three genes do not show any peptide, indicating that the operon might actually be composed of two separate operons.

same time, 10 (20%) of the 50 operons are found in *All Expressed* and *None Expressed* searches and are therefore considered validated by proteomic evidence.

Of particular interest are the “Split” cases, where a clear border is detected dividing the operon into two parts. 47 such cases were detected in total. For instance, as shown in Figure 7.6, one such case is the operon consisting of genes NT01PF1015 though NT01PF1017 in the *Pyrococcus furiosus* genome, as predicted by the microarray experiments [Tran07]. However, in mass spectrometry, no peptides are found for the last two proteins, while the first protein is supported by 12 peptides, giving it the binary signature “100”. Thus the proteomics data suggests splitting the operon between the genes NT01PF1015 and NT01PF1016. We note that the proteomics inference is consistent with the functional annotation of these genes, as the first gene is annotated as an ABC transporter, while the last two as “hypothetical protein”. This is an example of how proteomics data can provide valuable information to correct sequence or expression based operon predictions. Similarly, a pattern like “111110” or “000001” may represent a case when an operon might need to be shortened.

Figure 7.7 presents the case of an operon containing 6 genes, NT01PF1489 - NT01PF1494, in both the microarray-based and neural network-based operon predictions. The first three genes in the operon set (Figure 7.7) have 28, 4 and 8 peptides, respectively, identified through mass spectrometry, while the last three genes do not have any identified peptides. This is evidence in favour of the hypothesis that these genes might be belong to two distinct operons instead of one, the first one spanning NT01PF1489 - NT01PF1491 and the second one spanning NT01PF1492 - NT01PF1494. Indeed, looking the gene annotations reveals that the first three genes are annotated as kinase/ pyrophosphorylase, while the last three genes are annotated as putative genes with different functions, thus supporting the hypothesis.

7.4.2 Discussion

We have presented a simple framework for complementing sequence and expression based operon predictions with proteomic data derived from mass spectrometry. We have shown that protein expression can provide evidence for validating a predicted operon, or provide clues for correcting it by splitting or shortening at the ends. We note that this approach can be further refined if large-scale proteomic datasets are available under different experimental conditions, to detect sets of proteins that turn on or off together across conditions.

The proteomic evidence may be confounded by a few factors. Firstly, we assumed that proteins within the same operon are expressed together. While this is expected to be the case for mRNA transcripts, it is unclear if proteins within an operon are also expressed equally. However, since we only use qualitative (binarized) information about protein expression, our approach is expected to be robust against fluctuations in the level of protein expression as long as the proteins within an expressed operon produce even one detectable peptide. Inconsistencies between operon annotations and the proteomic data may also stem from secondary regulation of genes on the mRNA, protein degradation, or from the presence of non-protein-coding genes.

A second problem with using mass spectrometry is the possibility of not detecting an expressed protein. A low-abundance protein, or a short protein that has only a small number of proteotypic peptides [88] might not always get detected. However similar concerns also apply to microarrays and other experimental techniques.

Proteogenomic and comparative proteogenomic approaches [53,51] have shown the utility of mass spectrometry in improving genomic and proteomic annotations for newly sequenced organisms. With increasing availability of mass spectrometry data, and its efficacy in complementing sequence based operon predictions, we expect proteomics to become an important contributing technology for operon prediction and validation in the future.

7.5 Conclusion

The framework of proteogenomics developed in the previous chapters of this thesis was applied here for new biological discoveries, ranging from discovery of novel neuropeptides in human brain tissue, identification of mutations and rare post-translation modifications in bacterial genomes, and finding new structures of operons in archaea. These pilot studies demonstrate the potential of proteogenomics, and I believe that the coming years will see unfolding of many other novel applications of mass spectrometry that were not feasible before.

Chapter 7 includes parts from multiple manuscripts in preparation in which the disseration author is the first or the second author. “Evaluation of Alternative Neuropeptide Processing in Human and Bovine Dense-Core Secretory Granules by Mass Spectrometry-Based Neuropeptidomics. N. Gupta, S. Bark, W.D. Lu, L. Taupenot, D. O’Connor, P.A. Pevzner, V. Hook”, “MS-Operon: Using tandem mass spectrometry for operon prediction and validation. L. Wich and N. Gupta” and “Discovery of mutations and rare modifications using mass spectrometry. N. Gupta, R. D. Smith and P. A. Pevzner”.

References

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[2] S. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[3] P. Alves, R.J. Arnold, M.V. Novotny, P. Radivojac, J.P. Reilly, and H. Tang. Advancement in protein inference from shotgun proteomics using peptide detectability. *Pac Symp Biocomput*, pages 409–420, 2007.

[4] H. Antelmann, H. Tjalsma, B. Voigt, S. Ohlmeier, S. Bron, J.M. van Dijl, and M. Hecker. A Proteomic View on Genome-Based Signal Peptide Predictions. *Genome Res.*, 11:1484–1502, 2001.

[5] B.M. Austen and E.L. Smith. Action of staphylococcal proteinase on peptides of varying chain length and composition. *Biochem Biophys Res Commun*, 72:411–417, 1976.

[6] O. Bader, Y. Krauke, and B. Hube. Processing of predicted substrates of fungal Kex2 proteinases from Candida albicans, C. glabrata, Saccharomyces cerevisiae and Pichia pastoris. *BMC microbiology*, 8:116, 2008.

[7] N. Bandeira, K.R. Clauser, and P.A. Pevzner. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Molecular & Cellular Proteomics*, 6:1123–1134, 2007.

[8] N. Bandeira, V. Pham, P. Pevzner, D. Arnott, and J.R. Lill. Automated de novo protein sequencing of monoclonal antibodies. *Nature Biotechnology*, 26:1336–1338, 2008.

[9] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein identification by spectral networks analysis. *PNAS*, 104:6140–6145, 2007.

[10] A.J. Barrett, N.D. Rawlings, and J.F. Woessner. *Handbook of proteolytic enzymes*. Academic Press San Diego, 1998.

[11] S. Batzoglou, L. Pachter, J.P. Mesirov, B. Berger, and E.S. Lander. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Res*, 10:950–958, 2000.

[12] A. Ben-Bassat, K. Bauer, S.Y. Chang, K. Myambo, A. Boosman, and S. Chang. Processing of the initiation methionine from proteins: properties of the Escherichia coli methionine aminopeptidase and its gene structure. *J. Bacteriol.*, 169:751–757, 1987.

[13] J.D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340:783–795, Jul 2004.

[14] J. Bockhorst, Y. Qiu, J. Glasner, M. Liu, F. Blattner, and M. Craven. Predicting bacterial transcription units using sequence and expression data. *BIOINFORMATICS*, 19(1):i34–i43, 2003.

[15] K. Boonen, B. Landuyt, G. Baggerman, S.J. Husson, J. Huybrechts, and L. Schoofs. Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. *J Sep Sci*, 31:427–445, 2008.

[16] K.T. Boulware and P.S. Daugherty. Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proceedings of the National Academy of Sciences*, 103:7583–7588, 2006.

[17] R.A. Bradshaw, A.L. Burlingame, S. Carr, and R. Aebersold. Reporting Protein Identification Data: The next Generation of Guidelines. *Mol Cell Proteomics*, 5:787–8, 2006.

[18] RW Brouwer, OP Kuipers, and SA van Hijum. The relative value of operon predictions. *Briefings in bioinformatics*, 9(5):367–375, 2008.

[19] N.P. Brown, C. Sander, and P. Bork. Frame: detection of genomic sequencing errors. *Bioinformatics*, 14:367–371, 1998.

[20] B.J. Cargile, J.L. Bundy, and J.L. Stephenson Jr. Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of Proteome Research*, 3:1082–1085, 2004.

[21] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, and A. Nesvizhskii. The Need for Guidelines in Publication of Peptide and Protein Identification Data: Working Group On Publication Guidelines For Peptide And Protein Identification Data. *Mol Cell Proteomics*, 3:531, 2004.

[22] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, and J.D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31:3497–3500, 2003.

[23] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M.F. Lin, M. Kellis, K. Lindblad-Toh, and E.S. Lander. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, 104:19428–19433, 2007.

[24] K.R. Clauser, P. Baker, A.L. Burlingame, et al. Role of accurate mass measurement ($\pm$10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem*, 71:2871–2882, 1999.

[25] J.M. Claverie. Detecting frame shifts by amino acid sequence comparison. *J Mol Biol*, 234:1140–57, 1993.

[26] R. Craig and R.C. Beavis. TANDEM: matching proteins with tandem mass-spectra. *Bioinformatics*, 20:1466–1467, 2004.

[27] W.J. Craigen, R.G. Cook, W.P. Tate, and C.T. Caskey. Bacterial Peptide Chain Release Factors: Conserved Primary Structure and Possible Frameshift Regulation of Release Factor 2. *Proceedings of the National Academy of Sciences*, 82:3616–3620, 1985.

[28] G.E. Crooks, G. Hon, J. Chandonia, and S.E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14:1188–1190, 2004.

[29] N. Daraselia, D. Dernovoy, Y. Tian, M. Borodovsky, R. Tatusov, and T. Tatusova. Reannotation of Shewanella oneidensis Genome. *OMICS: A Journal of Integrative Biology*, 7:171–176, 2003.

[30] P.A. Demirev, J.S. Lin, F.J. Pineda, and C. Fenselau. Bioinformatics and mass spectrometry for microorganism identification: proteome-wide post-translational modifications and database search algorithms for characterization of intact H. pylori. *Analytical Chemistry*, 73:4566–4573, 2001.

[31] D. Deperthes. Phage display substrate: a blind method for determining protease specificity. *Biol. Chem*, 383:1107–1112, 2002.

[32] F. Desiere, E. Deutsch, A. Nesvizhskii, P. Mallick, N. King, J. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology*, 6:R9, 2004.

[33] R.C. Edgar and O. Journals. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

[34] D.A. Elias, M.E. Monroe, R.D. Smith, J.K. Fredrickson, and M.S. Lipton. Confirmation of the expression of a large set of conserved hypothetical proteins in Shewanella oneidensis MR-1. *Journal of Microbiological Methods*, 66:223–233, 2006.

[35] Dwayne A Elias, Matthew E Monroe, Matthew J Marshall, Margaret F Romine, Alexander S Belieav, James K Fredrickson, Gordon A Anderson, Richard D Smith, and Mary S Lipton. Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. *Proteomics*, 5(12):3120–3130, Aug 2005.

[36] J.E. Elias and S.P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4:207–214, 2007.

[37] M. Enoksson, J. Li, M.M. Ivancic, J.C. Timmer, E. Wildfang, A. Eroshkin, G.S. Salvesen, and W.A. Tao. Identification of Proteolytic Cleavage Sites by Quantitative Proteomics. *J. Proteome Res*, 6:2850–2858, 2007.

[38] M.D. Ermolaeva, O. White, and S.L. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Research*, 29:1216–1221, 2001.

[39] M. Falth, K. Skold, M. Norrman, M. Svensson, D. Fenyo, and P.E. Andren. SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell Proteomics*, 5:998–1005, 2006.

[40] P.J. Farabaugh. Programmed translational frameshifting. *Microbiology and Molecular Biology Reviews*, 60:103–134, 1996.

[41] J. Feng, D.Q. Naiman, and B. Cooper. Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Anal. Chem*, 79:3901–3911, 2007.

[42] 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*, 7:R35, 2006.

[43] G.A. Fichant and Y. Quentin. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res*, 23:2900–2908, 1995.

[44] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. *Science*, 269:496–498, 1995.

[45] 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*, 7:113–122, 2008.

[46] F. Frottin, A. Martinez, P. Peynot, S. Mitra, R.C. Holz, e C. Giglion, and T. Meinnel. The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics*, 5:2336–2349, 2006.

[47] T. Georgiou, Y.T.N. Yu, S. Ekunwe, MJ Buttner, A.M. Zuurmond, B. Kraal, C. Kleanthous, and L. Snyder. Specific peptide-activated proteolytic cleavage of Escherichia coli elongation factor Tu. *Proceedings of the National Academy of Sciences*, 95:2891–2895, 1998.

[48] E. Gross and B. Witkop. Selective cleavage of the methionyl peptide bonds in ribonuclease with cyanogen bromide1. *Journal of the American Chemical Society*, 83:1510–1511, 1961.

[49] L.J. Guibas and A.M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Comb. Theory, Ser. A*, 30:183–208, 1981.

[50] N. Gupta, S.J. Bark, W.D. Lu, L. Taupenot, D.T. O'Connor, P.A. Pevzner, and V. Hook. Evaluation of Alternative Neuropeptide Processing in Human and Bovine Dense-Core Secretory Granules by Mass Spectrometry-Based Neuropeptidomics. *Submitted.*

[51] N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, 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.*, 18:1133–1142, 2008.

[52] N. Gupta and P.A. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *(under revision).*

[53] N. Gupta, S. Tanner, N. Jaitly, J. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. Smith, and P. Pevzner. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res*, 17:1362–1377, 2007.

[54] J.L. Harris, B.J. Backes, F. Leonetti, S. Mahrus, J.A. Ellman, and C.S. Craik. Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc Natl Acad Sci US A*, 97:7754–9, 2000.

[55] J.F. Heidelberg et al. Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. *Nature Biotechnology*, 20:1118–1123, 2002.

[56] R. Higdon and E. Kolker. A predictive model for identifying proteins by a single peptide match. *Bioinformatics*, 23:277–280, 2007.

[57] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*, 32:375–379, 2004.

[58] P H Hirel, M J Schmitter, P Dessen, G Fayat, and S Blanquet. Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci U S A*, 86(21):8247–8251, Nov 1989.

[59] J. Houmard and G.R. Drapeau. Staphylococcal Protease: A Proteolytic Enzyme Specific for Glutamoyl Bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 69:3506–3509, 1972.

[60] WS Hu, RY Wang, JW Shih, and SC Lo. Identification of a putative infC-rpmI-rplT operon flanked by long inverted repeats in Mycoplasma fermentans (incognitus strain). *Gene*, 127(1):79–85, 1993.

[61] D.F. Hunt, J.R. Yates, J. Shabanowitz, S. Winston, and C.R. Hauer. Protein Sequencing by Tandem Mass Spectrometry. *Proceedings of the National Academy of Sciences*, 83:6233–6237, 1986.

[62] E. Hunyadi-Gulyas and K. Medzihradszky. Factors that contribute to the complexity of protein digests. *DDT: targets - mass spectrometry in proteomics supplement*, 3(2):S3–S10, 2004.

[63] Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J.W. Smith, A.L. Osterman, and A. Godzik. CutDB: a proteolytic event database. *Nucleic Acids Research*, 35:D546, 2007.

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

[65] 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*, 14:1447–1461, 2004.

[66] L.J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, and P. Bork. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36:250, 2008.

[67] L. Kall, J.D. Canterbury, J. Weston, W.S. Noble, and M.J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–926, 2007.

[68] L. Kall, 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.*, 7:29–34, 2008.

[69] 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*, 2005:128, 2005.

[70] B. Keil. Proteolysis Data Bank: specificity of alpha-chymotrypsin from computation of protein cleavages. *Protein Seq Data Anal*, 1:13–20, 1987.

[71] B. Keil. *Specificity of Proteolysis*. Springer-Verlag Berlin, Germany, 1992.

[72] A. Keller, A.I. 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*, 74:5383–5392, Oct 2002.

[73] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.

[74] S. Kim, N. Gupta, N. Banderia, and P.A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8:53–69, 2009.

[75] S. Kim, N. Gupta, and P.A. Pevzner. Spectral Probabilities and Generating Functions of Tandem Mass Spectra: a Strike Against Decoy Databases. *J. Proteome Res.*, 7:3354–3363, 2008.

[76] E. Kolker, A.F. Picone, M.Y. Galperin, M.F. Romine, R. Higdon, K.S. Makarova, N. Kolker, G.A. Anderson, X. Qiu, K.J. Auberry, et al. Global profiling of Shewanella oneidensis MR-1: Expression of hypothetical genes and improved functional annotations. *Proceedings of the National Academy of Sciences*, 102:2099–2104, 2005.

[77] J A Kowalak and K A Walsh. Beta-methylthio-aspartic acid: identification of a novel posttranslational modification in ribosomal protein S12 from Escherichia coli. *Protein Sci*, 5(8):1625–1632, Aug 1996.

[78] B Kuster, P Mortensen, J S Andersen, and M Mann. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, 1(5):641–650, May 2001.

[79] B. Kuster, M. Schirle, P. Mallick, and R. Aebersold. Scoring Proteomes With Proteotypic Peptide Probes. *Nature Reviews Molecular Cell Biology*, 6:577–583, 2005.

[80] L. Li and J.V. Sweedler. Peptides in the Brain: Mass Spectrometry–Based Measurement Approaches and Challenges. *Annu. Rev. Anal. Chem.*, 1:451–483, 2008.

[81] Y.F. Li, R.J. Arnold, Y. Li, P. Radivojac, Q. Sheng, and H. Tang. A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. *RECOMB*, pages 167–180, 2008.

[82] Y.D. Liao, J.C Jeng, C.F. Wang, S.C. Wang, and S.T. Chang. Removal of N-terminal methionine from recombinant proteins by engineered E. coli methionine aminopeptidase. *Protein Science*, 13:1802–1810, 2004.

[83] AJ Link, K. Robison, and GM Church. Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. *Electrophoresis*, 18(8):1259–313, 1997.

[84] S. Liu, G.T. Milne, J.G. Kuremsky, G.R. Fink, and S.H. Leppla. Identification of the proteins required for biosynthesis of diphthamide, the target of bacterial ADP-ribosylating toxins on translation elongation factor 2. *Mol. Cell. Biol.*, 24:9487–9497, 2004.

[85] D. Liveris, J.J. Schwartz, R. Geertman, and I. Schwartz. Molecular cloning and sequencing of encoding translation initiation factor enterobacterial species. *FEMS Microbiology Letters*, 112:211–216, 1993.

[86] P. Lu, C. Vogel, R. Wang, X. Yao, and E.M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, 25:117–124, 2007.

[87] M.J. MacCoss, C.C. Wu, H. Liu, R. Sadygov, and J.R. Yates III. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem*, 75:6912–21, 2003.

[88] P. Mallick, M. Schirle, SS Chen, MR Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol*, 25:125–131, 2007.

[89] N.P. Manes, J.K. Gustin, J. Rue, H.M. Mottaz, S.O. Purvine, A.D. Norbeck, M.E. Monroe, J.S.D. Zimmer, T.O. Metz, J.N. Adkins, et al. Targeted Protein Degradation by Salmonella under Phagosome-mimicking Culture Conditions Investigated Using Comparative Peptidomics. *Molecular and Cellular Proteomics*, 6:717–727, 2007.

[90] M. Mann and A. Pandey. Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem Sci*, 26(1):54–61, 2001.

[91] C. Masselon, L. Pasa-Tolic, N. Tolic, G.A. Anderson, B. Bogdanov, A.N. Vilkov, Y. Shen, R. Zhao, W.J. Qian, M.S. Lipton, et al. Targeted comparative proteomics by liquid chromatography-tandem Fourier ion cyclotron resonance mass spectrometry. *Anal. Chem*, 77:400–406, 2005.

[92] C. Medigue, M. Rose, A. Viari, and A. Danchin. Detecting and Analyzing DNA Sequencing Errors: Toward a Higher Quality of the Bacillus subtilis Genome Sequence. *Genome Res*, 9:1116–1127, 1999.

[93] J.M. Moehring, T.J. Moehring, and D.E. Danley. Posttranslational modification of elongation factor 2 in diphtheria-toxin-resistant mutants of CHO-K1 cells. *Proc. Natl. Acad. Sci. U.S.A.*, 77:1010–1014, 1980.

[94] R.E. Moore, M.K. Young, and T.D. Lee. Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13:378–386, 2002.

[95] K.H. Nealson, A. Belz, and B. McKee. Breathing metals as a way of life: geobiology in action. *Antonie Van Leeuwenhoek*, 81:215–222, 2002.

[96] A.I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics*, 4:1419–1440, 2005.

[97] A.I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75:4646–4658, 2003.

[98] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, 10:1–6, 1997.

[99] M.L. Nielsen, M.M. Savitski, and R.A. Zubarev. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics*, 5:2384–2391, 2006.

[100] J.V. Olsen, S. Ong, and M. Mann. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics*, 3:608–614, 2004.

[101] K.C. Olson, J. Fenno, N. Lin, R.N. Harkins, C. Snider, WH Kohr, M.J. Ross, D. Fodge, G. Prender, and N. Stebbing. Purified human growth hormone from E. coli is biologically active. *Nature*, 293:408–411, 1981.

[102] G.S. Omenn, D.J. States, M. Adamski, T.W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B.B. Haab, R.J. Simpson, J.S. Eddes, et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, 5:3226–3245, 2005.

[103] G. Oshiro, L.M. Wodicka, M.P. Washburn, J.R. Yates III, D.J. Lockhart, and E.A. Winzeler. Parallel Identification of New Genes in Saccharomyces cerevisiae. *Genome Res*, 12(8):1210–1220, 2002.

[104] Mark Paetzel, Andrew Karla, Natalie C J Strynadka, and Ross E Dalbey. Signal peptidases. *Chem Rev*, 102(12):4549–4580, Dec 2002.

[105] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[106] C.L. Pon, M. Brombach, S. Thamm, and C.O. Gualerzi. Cloning and characterization of a gene cluster from Bacillus stearothermophilus comprising infC, rpmI and rplT. *Molecular Genetics and Genomics*, 218(2):355–357, 1989.

[107] J. Posfai and R.J. Roberts. Finding Errors in DNA Sequences. *Proceedings of the National Academy of Sciences*, 89:4698–4702, 1992.

[108] M.N. Price, K.H. Huang, E.J. Alm, and A.P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research*, 33:880–892, 2005.

[109] S. Purvine, A.F. Picone, and E. Kolker. Standard Mixtures for Proteome Studies. *Omics A Journal of Integrative Biology*, 8:79–92, 2004.

[110] T.A. Rano, T. Timkey, E.P. Peterson, J. Rotonda, D.W. Nicholson, J.W. Becker, K.T. Chapman, and N.A. Thornberry. A combinatorial approach for determining protease specificities: application to interleukin-1 converting enzyme (ICE). *Chem. Biol*, 4:149–155, 1997.

[111] M. Remm, C.E. Storm, and E.L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*, 314:1041–1052, 2001.

[112] J. Rodriguez, N. Gupta, R.D. Smith, and P.A. Pevzner. Does trypsin cut before proline? *J Proteome Res*, 7:300–305, 2008.

[113] Margaret F Romine, Dwayne A Elias, Matthew E Monroe, Kenneth Auberry, Ruihua Fang, Jim K Fredrickson, Gordon A Anderson, Richard D Smith, and Mary S Lipton. Validation of Shewanella oneidensis MR-1 small proteins by AMT tag-based proteome analysis. *OMICS*, 8(3):239–254, Fall 2004.

[114] C. Sacerdot, G. Fayat, P. Dessen, M. Springer, JA Plumbridge, M. Grunberg-Manago, and S. Blanquet. Sequence of a 1.26-kb DNA fragment containing the structural gene for E. coli initiation factor IF3: presence of an AUU initiator codon. *EMBO J*, 1(3):311–315, 1982.

[115] H. Salgado, G. Moreno-Hagelsieb, TF Smith, and J. Collado-Vides. Operons in Escherichia coli: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences, USA*, 97(12):6652–6657, 2000.

[116] A. Savidor, R.S. Donahoo, O. Hurtado-Gonzales, N.C. Verberkmoes, M.B. Shah, K.H. Lamour, and W.H. McDonald. Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation. *Journal of proteome research*, 5:3048–3058, 2006.

[117] M.M. Savitski, M.L. Nielsen, and R.A. Zubarev. Modificomb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics*, 5:935–948, 2006.

[118] V. Schellenberger, K. Braune, H.J. Hofmann, and H.D. Jakubke. The specificity of chymotrypsin. *Eur. J. Biochem*, 199:623–636, 1991.

[119] O. Schilling and C.M. Overall. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature Biotechnology*, 26:685–694, 2008.

[120] G J Schoenhals, M Kihara, and R M Macnab. Translation of the flagellar gene fliO of Salmonella typhimurium from putative tandem starts. *J Bacteriol*, 180(11):2936–2942, Jun 1998.

[121] Y. Shen, K.K. Hixson, N. Tolic?, D.G. Camp, S.O. Purvine, R.J. Moore, and R.D. Smith. Mass Spectrometry Analysis of Proteome-Wide Proteolytic Post-Translational Degradation of Proteins. *Analytical Chemistry*, 80:5819–5828, 2008.

[122] S.B. Sorensen, T.L. Sorensen, and K. Breddam. Fragmentation of proteins by S. aureus strain V 8 protease: Ammonium bicarbonate strongly inhibits the enzyme but does not improve the selectivity for glutamic acid. *FEBS Letters*, 294:195–197, 1991.

[123] J K Sussman, E L Simons, and R W Simons. Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol Microbiol*, 21(2):347–360, Jul 1996.

[124] H. Tang, R.J. Arnold, P. Alves, Z. Xun, D.E. Clemmer, M.V. Novotny, J.P. Reilly, and P. Rejivojac. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, 22:e481–e488, 2006.

[125] S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigo, S.P. Briggs, and V. Bafna. Improving gene annotation using peptide mass spectrometry. *Genome Res*, 17:231–239, 2007.

[126] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005.

[127] A.N. Tegge, B.R. Southey, J.V. Sweedler, and S.L. Rodriguez-Zas. Comparative analysis of neuropeptide cleavage sites in human, mouse, rat, and cattle. *Mammalian Genome*, 19:106–120, 2008.

[128] M.R. Thompson, D.K. Thompson, and R.L. Hettich. Systematic assessment of the benefits and caveats in mining microbial post-translational modifications from shotgun proteomic data: the response of Shewanella oneidensis to chromate exposure. *J. Proteome Res.*, 7:648–658, 2008.

[129] R. Thomson, T.C. Hodgman, Z.R. Yang, and A.K. Doyle. Characterizing proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics*, 19:1741–1747, 2003.

[130] T.A. Thornberry, N.A.and Rano, E.P. Peterson, D.M. Rasper, T. Timkey, M. Garcia-Calvo, V.M. Houtzager, P.A. Nordstrom, S. Roy, J.P. Vaillancourt, et al. A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem*, 272:17907–11, 1997.

[131] J.C. Timmer, M. Enoksson, E. Wildfang, W. Zhu, Y. Igarashi, J.B. Denault, Y. Ma, B. Dummitt, Y.H. Chang, A.E. Mast, A. Eroshkin, J. Smith, W.A. Tao, and G.S. Salvesen. Profiling constitutive proteolytic events in vivo. *Biochem. J*, 407:41–48, 2007.

[132] J.W. Tobias, T.E. Shrader, G. Rocap, and A. Varshavsky. The N-end rule in bacteria. *Science*, 254:1374–1377, 1991.

[133] T.T. Tran, P. Dam, Z. Su, F.L.I.I. Poole, M.W.W. Adams, G.T. Zhou, and Y. Xu. Operon prediction in Pyrococcus furiosus. *Nucleic Acids Research*, 35:11–20, 2007.

[134] S.A. Trauger, E. Kalisak, J. Kalisiak, H. Morita, M.V. Weinberg, A.L. Menon, F.L. Poole Ii, M.W.W. Adams, and G. Siuzdak. Correlating the transcriptome, proteome, and metabolome in the environmental adaptation of a hyperthermophile. *Journal of Proteome Research*, 7:1027–1035, 2008.

[135] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P.A. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. *Nature Biotechnology*, 23:1562–2567, 2005.

[136] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P.A. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. *Nature Biotechnology*, 23:1562–2567, 2005.

[137] C. Tu, T. Tzeng, and JA Bruenn. Ribosomal Movement Impeded at a Pseudoknot Required for Frameshifting. *Proceedings of the National Academy of Sciences*, 89:8636–8640, 1992.

[138] B.E. Turk, L.L. Huang, E.T. Piro, and L.C. Cantley. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nature Biotechnology*, 19:661–667, 2001.

[139] B.G. Van Ness, J.B. Howard, and J.W. Bodley. ADP-ribosylation of elongation factor 2 by diphtheria toxin. NMR spectra and proposed structures of ribosyl-diphthamide and its hydrolysis products. *J. Biol. Chem.*, 255:10710–10716, 1980.

[140] 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*, 15:1118–1126, 2005.

[141] M.P. Washburn, D. Wolters, and J.R. Yates III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19:242–247, 2001.

[142] V.C. Wasinger and I. Humphery-Smith. Small genes/gene-products in Escherichia coli K-12. *FEMS Microbiol Lett.*, 169:375–382, 1998.

[143] D.B. Weatherly, J.A. Atwood, T.A. Minning, C. Cavola, R.L. Tarleton, and R. Orlando. A Heuristic Method for Assigning a False-discovery Rate for Protein Identifications from Mascot Database Search Results. *Molecular & Cellular Proteomics*, 4:762–772, 2005.

[144] P.A. Wilmarth, S. Tanner, S. Dasari, S.R. Nagalla, M.A. Riviere, V. Bafna, P.A. Pevzner, and L.L. David. Age-Related Changes in Human Crystallins Determined from Comparative Analysis of Post-translational Modifications in Young and Aged Lens: Does Deamidation Contribute to Crystallin Insolubility? *J. Proteome Res*, 5:2554–2566, 2006.

[145] C.C. Wu and J.R. Yates 3rd. The application of mass spectrometry to membrane proteomics. *Nature Biotechnology*, 21:262–7, 2003.

[146] X. Xie, J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434:338–345, 2005.

[147] C. Yang, D.A. Rodionov, X. Li, O.N. Laikova, M.S. Gelfand, O.P. Zagnitko, M.F. Romine, A.Y. Obraztsova, K.H. Nealson, and A.L. Osterman. Comparative Genomics and Experimental Characterization of N-Acetylglucosamine Utilization Pathway of Shewanella oneidensis. *Journal of Biological Chemistry*, 281:29872–29885, 2006.

[148] J.R. Yates, J.K. Eng, and A.L. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67(18):3202–3210, Sep 1995.

[149] Y. Zheng, J.D. Szustakowski, L. Fortnow, R.J. Roberts, and S. Kasif. Computational Identification of Operons in Microbial Genomes. *Genome Research*, 12(8):1221, 2002.